Università degli Studi di Padova

Dipartimento di Biologia

SCUOLA DI DOTTORATO DI RICERCA IN: BIOSCIENZE E BIOTECNOLOGIE

INDIRIZZO: GENETICA E BIOLOGIA MOLECOLARE DELLO SVILUPPO

CICLO XXIV

THE DYNAMIC CHANGES OF EXPRESSION SIGNATURES IN THE LARVAL DEVELOPMENT OF *MYTILUS GALLOPRONVINCIALIS*

Direttore della Scuola: Ch.mo Prof. Giuseppe Zanotti

Coordinatore d'indirizzo: Ch.mo Prof. Paolo Bonaldo

Supervisore: Ch.mo Prof. Gerolamo Lanfranchi

Dottorando: Alberto Biscontin

INDEX

Abstract 1

Riassunto 3

1 Introduction 5

1.1 Mediterranean mussel (*Mytilus galloprovincialis*) 5
1.1.1 Origin and distribution 5
1.1.2 Morphological aspects 5
1.1.3 Ecological aspects of *M. galloprovincialis* 9
1.1.4 Reproductive cycle 9
1.1.5 Larval development 10
1.1.6 Ecological and economical relevance of *Mytilus galloprovincialis* 12
1.1.7 Genomic, transcriptomic, and proteomic information about *M. galloprovincialis* 13

1.2 Analysis of the transcriptome 14
1.2.1 Comparison between microarray and RNA sequencing technology 14
1.2.2 Second-generation massively sequencing platforms 17
1.2.3 Applications of next generation platforms 20
1.2.4 Single molecule sequencing platforms 21
1.2.5 Non-optic sequencing 22

2 Aims 25

3 Materials and Methods 27

3.1 Primers table

3.2 Mating experiments 28
3.2.1 Facilities 28
3.2.2 Spawning 28
3.2.3 Rearing 29
3.2.4 Sampling 30

3.3 Sample processing 31
3.3.1 Total RNA extraction and quality control 31
3.3.2 Bouin fixation 31
3.3.3 PFA fixation 32

3.4 Construction of the 3'-end cDNA libraries 32
3.4.1 The first-strand synthesis of cDNA using SMART technology 32
3.4.2 cDNA amplification by PCR 32
3.4.3 Fragmentation of double-strand cDNA by nebulization 33
3.4.4 The 3'-end selection and poly(A) processing 33

3.5 3'-end cDNA libraries sequencing 34

3.6 Data analysis **34**
3.6.1 Alignments between 454 reads and EST sequences deposited on MytiBase 34
3.6.2 Assembling of 454 reads 35
3.6.3 Annotation process 36
3.6.4 Definition of relative abundance of transcripts 36
3.6.5 Clustering analysis 36
3.6.6 Identification of differentially expressed genes 36
3.6.7 Functional annotation 37

3.7 Quantitaltive Real Time PCR (qRT-PCR) **37**

3.8 *In situ* hybridization **38**
3.8.1 *In situ* probe design 38
3.8.2 Whole mount in situ hybridization 39
3.8.3 Mounting 40

3.9 Recovery of full length **40**

4 Results **43**

4.1 Experimental design 43

4.2 Collection of mussel specific developmental stage 44

4.3 Total RNA extraction and quality control 45

4.4 Setting up of the 3'-end cDNA libraries **46**
4.4.1 First strand cDNA synthesis 47
4.4.2 PCR amplification of the ssDNA 49
4.4.3 Fragmentation of double-strand cDNA 49
4.4.4 Selection of 3'-end cDNA fragments 51
4.4.5 Recovery of 3'-end cDNA fragments from magnetic beads 51
4.4.6 Quality and quantity analysis of cDNA libraries 53

4.5 Next generation sequencing of larval specific cDNA libraries **54**

4.6 Analysis of 454 sequencing data to verify the library protocol **55**

4.7 454 reads assembling and construction of a mussel development transcript catalogue **57**
4.7.1 Assembling process 57
4.7.2 Has 454 sequencing increased the transcriptome knowledge of *M. galloprovincialis*? 59
4.7.3 Annotation process 60

4.8 Defining larval development expression signature **61**
4.8.1 454 reads counting and cluster analysis 61
4.8.2 Determination of stage-specific expressed genes in *M. galloprovincialis* development 63

4.9 Validation of differentially expressed genes by quantitative real-time PCR (qRT-PCR) **64**

4.10 Gene Ontology analysis **65**
4.11 Functional gene characterization **67**
4.12 Embryo library analysis **69**

5 Discussion **71**

6 Conclusions **79**

7 References **81**

Appendix 1 87

Appendix 2 92

Acknowledgments 93

Abstract

THE DYNAMIC CHANGES OF EXPRESSION SIGNATURES IN THE LARVAL DEVELOPMENT OF *MYTILUS GALLOPRONVINCIALIS*

Bivalves are some of the most studied marine organisms for ecological importance as bio-indicators and economic value in shellfish farming. They have a planktonic life stage, spending about 4 weeks as free-swimming larvae in the water column and probably this is an evolutive advantage making the most common marine molluscs. A variety of physical, biological, and genetic factors influence the initial larvae pelagic dispersion and the subsequent soil settlement. During larval stages bivalves are characterized by a very fast growth, and by morphological and metabolic changes pointed to the accomplishment of metamorphosis into adult organism. These radical modifications are probably due to the activation of different gene expression patterns associated with each developmental stage.

The aim of my PhD project was to define the transcriptional signatures of 6 larval stages in *Mytilus galloprovincialis* to provide a comprehensive overview on the biological processes underlying the larval development.

In 2009 we performed the stimulation of bivalves mating to collect samples of fertilized eggs and each larval stage. Total RNA have been extracted and each sample was fixed 4% paraformaldehyde for *in situ* hybridization.

We set up seven stage-specific 3'-end cDNA libraries in order: i) to discover specifically expressed transcripts in larval stages, and ii) to define the first transcriptional signature of larval development in *M. galloprovincialis*, calculating the 3'-EST frequency at each stage. Using pyrosequencing technology (Roche 454 Titanium), we have generated a total of 751,872 high quality reads characterized by a median length of 315 bp. The assembling process (Newbler 2.5.1) of pyrosequencing reads and ESTs, previously produced by our lab (Mytibase), resulted in 14,364 unique sequences/putative transcripts and 55,493 singletons. We counted the number of reads that align with each unique sequence at each stage-specific library defining the expression pattern of each stage.

After the annotation process we were able to estimate a gene discovery rate of about 45% rising to 60% in the early larval stages characterized by a lot of unknown genes. Quantitative real time PCR analysis was performed to quantify and validate the expression profile of some genes on samples collected during a new mussel mating experiment, performed at the beginning of 2011. On the basis of the high correlation values between the qRT-PCR and sequencing data, we can consider the counts reliable. A negative binomial test was performed to identify differentially expressed transcripts between at least two stages. In particular the 25% of genes results to be differentially expressed and we observed the main differences comparing early larval stages, including trocophora and D-larva, and late larval stages, including umbo stage, pediveliger, and metamorphic larva. Differentially expressed genes showing similarity with known genes or proteins were grouped into functional categories according to Gene Ontology in order to identify the biological categories involved in development such as organ development, growth control, biomineralization, and byssus secretion.

Finally we are performing whole mount *in situ* hybridization of the most differentially expressed unknown genes to start their functional characterization defining their morphological localization.

We defined the first transcriptional signatures of larval stages in *M. galloprovincialis* in order to better understand the molecular mechanisms involved in the process of development from larva to the adult.

Riassunto

STUDIO DEI PROFILI TRASCRIZIONALI ASSOCIATI A CIASCUN STADIO DI SVILUPPO LARVALE IN *MYTILUS GALLOPROVINCIALIS*

I bivalvi sono la classe di organismi marini più studiata, sia per il loro ruolo ecologico di bioindicatori del grado di inquinamento costiero, sia per la loro rilevanza economica nel settore alimentare. Il loro sviluppo larvale è caratterizzato da una rapida crescita e da profondi cambiamenti morfologici e metabolici volti al raggiungimento dello stadio adulto. I radicali mutamenti a cui l'organismo va incontro sono riconducibili all'espressione coordinata di specifici geni.

Lo scopo del mio progetto di dottorato è stato quello di definire i profili trascrizionali associati a ciascuno dei 6 stadi di sviluppo larvale di *Mytilus galloprovincialis* per far luce sui processi biologici alla base dell sviluppo larvale.

Nel 2009 abbiamo riprodotto in laboratorio il naturale processo di fecondazione tra 8 femmine e 20 maschi inducendo l'emissione dei gameti. Abbiamo quindi raccolto campioni di RNA totale dell'uovo fecondato e di ciascuno stadio larvale e abbiamo inoltre conservato in paraformaldeide alcune larve per effettuare esperimenti di ibridazione in situ.

A partire dall'RNA totale estratto abbiamo allestito sette librerie stadio specifiche costituite da frammenti 3' terminali di cDNA. Queste sono state sequenziale con le moderne tecnologie di sequenziamento massivo così da ottenere sia le sequenza di trascritti espressi durante gli stadi larvali, sia una stima del loro profilo di espressione nei vari stadi larvali calcolando la loro frequenza di rappresentazione in ciascuna libreria. Per il sequenziamento ci siamo avvalsi della tecnologia 454 (Roche) che ci ha permesso di ottenere 751,872 *reads* caratterizzate da una lunghezza mediana di 315 bp. Le reads ottenute sono state assemblate con sequenze prodotte precedntemente dal nostro laboratorio e depositate su MytiBase ottenendo 14,364 sequenze *consensus* e 55,493 *singletons.* Quindi abbiamo calcolato il numero di reads che si allineano con ciascun trascritto in ciascuna libreria stadio specifica così da ottenere una stima dei profili trascrizionali associati a ciascuno stadio di sviluppo larvale.

Sfruttando un algoritmo da noi sviluppato siamo riusciti ad annotare il 55% dei trascritti. Per verificare l'attendibilità delle conte come mezzo per stimare l'espressione genica abbiamo eseguito una serie di esperimenti di PCR quantitativa su un nuovo campionamento biologico eseguito all'inizio del 2011. Sulla base dell'elevata correlazione tra i dati di espressione stimati con le due tecnologie indipendenti possiamo giudicare affidabile il nostro approccio. Tramite un test binomiale negativo abbiamo determinato che il 25% dei trascritti risulta differenzialmente espresso in almeno due stadi di sviluppo. Inoltre le maggiori differenze di espressione si rilevano confrontando gli stadi precoci (troco fora e larva D) con quelli tardivi (larva umbonata, pediveliger e larva metamorfica). Mediante analisi di *Gene Ontology* abbiamo individuato i processi biologici maggiormente implicati nello sviluppo larvale quali ad esempio il controllo della crescita, lo sviluppo degli organi e i processi di biomineralizzazone e sintesi del bisso.

Infine stiamo eseguendo degli esperimenti di ibridazione *in situ* per incominciare la caratterizzazione di alcuni dei geni, differenzialmente espressi ma non annotati, definendo la loro localizzazione.

Concludendo, abbiamo ottenuto la prima stima dei profili trascrizionali associati a ciascuno stadio larvale di M. galloprovincialis per comprendere i meccanismi molecolari su cui fonda il processo di sviluppo larvale.

1 Introduction

1.1 Mediterranean mussel (*Mytilus galloprovincialis*)

1.1.1 Origin and distribution

Mytilus galloprovincialis (Lamark 1819), also known as the bay mussel or blue mussel, is a species of bivalve marine mollusc of the Mytilidae family (Phylum Mollusca; Class Bivalvia). *Mytilus* genus dates back to the Jurassic, between 200 and 300 millions years ago, but the species *M. galloprovincialis* appeared in Mediterranean sea during the last 2 millions years. The warmer temperatures of the Mediterranean sea and the lowering temperatures of oceans during the quaternary glaciations may have favoured the differentiation of Mediterranean mussel from *Mytilus edulis*, that is the ancestor species. (Skibinski DO; 1979). Now *M. galloprovincialis* is commonly found in the Mediterranean Sea, Black Sea, but also along the Atlantic coasts of France, Britain and Ireland, as well as in South Africa, California, China, Japan and the south coasts of the Australia. This world wide distribution is mainly due to the accidental transportation in the ballast waters of shipping vessels since the early 19th Century.

1.1.2 Morphological aspects

The features that distinguish Bivalves from other molluscs are two. The first, and most important difference, is the presence of two mantle lobes that enclose the internal organs from both sides. Two shell valves are secreted by the mantle and hinged together to protect the mollusc soft body. The second difference is represented by the lateral compression, due to the development of the shell, that has led mouth to lose its role of catching food that was shifted to gills. So, gills have become one of the most efficient systems of ciliary feeding in the Animal Kingdom. (Gosling E; 2003). Now we briefly describe the main morphological structures of *M. galloprovincialis* that are represented in Figure 1.1.

Shell. Under optimal conditions *M. galloprovincialis* shell length can ranges from 10.0 to 13.0 cm, whereas in the intertidal zone mussels may measure as little as 2.0–3.0 cm. In mussels the two asymmetric shell valves are linked together by the hinge ligament. Adult shell is a structure of calcium carbonate that is deposited on a cell-free proteinaceous matrix and consists of three layers: a thin outer proteinaceous covering, named periostracum, often eroded by mechanical abrasion, fouling organisms, parasites or disease; a middle prismatic layer of aragonite or calcite, depending on the local

environment; and an inner calcareous layer (named nacreous layer), that is iridescent mother-of-pearl. Mantle edge secretes the ionic precursors together with the components of the organic matrix. From a molecular point of view, matrix consists of a core of parallel sheets of β-chitin that are surrounded by a silk-like gel composed of several protein families that stabilize amorphous calcium carbonate and induce specific crystallization into aragonite or calcite. Chemical and physical structural properties of the shell are regulated by the protein composition of the silk-like protein gel (Weiss IM; 2006).

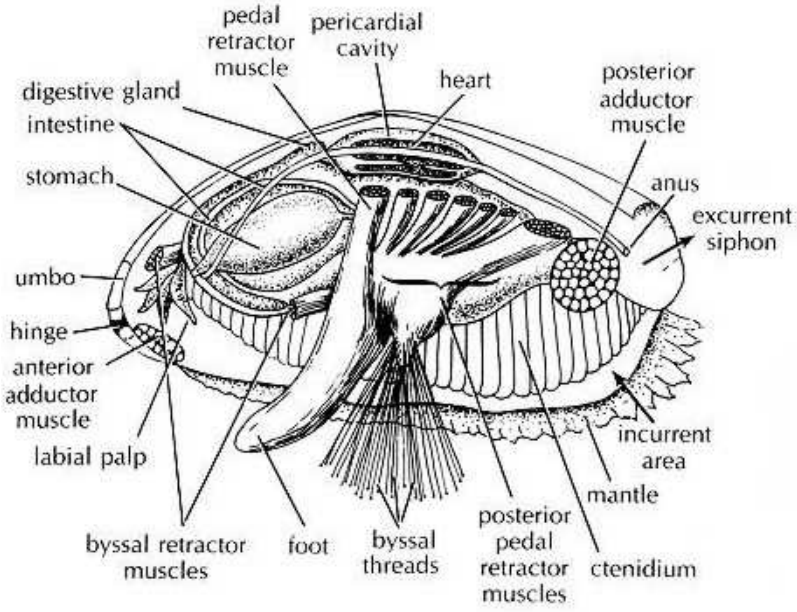

Figure 1.1 : Mussel morphology. Longitudinal section showing the major organs and morphological structures of *Mytilus galloprovincialis.*

Muscles. The opening and closing of the valves are controlled by two muscles that directly bond to the shell inner surface: the anterior and posterior adductor muscles. Moreover, these muscles have the ability, called catch state, to passively maintain force and resist stretch after the cessation of stimulation (Bardales JR; 2011) maintaining shell closed for long periods. Also the anterior and posterior retractor muscles, responsible for the foot-byssus movements, originate from the shell.

Mantle. In bivalves the mantle consists of two lobes of tissue which completely enclose the soft body. The mantle has multiple functions: it forms a capacious cavity that contains the gills and it is covered of cilia that play an important role in catching food. Futhermore, mantle is the site of gametogenesis and for the storage of nutrient reserves,

especially glycogen. Reserves are laid down in summer and are utilised in autumn and winter in the formation of gametes (Zwaan A, 1992). However, shell formation is the primary function and takes place along the mantle edge. Margins consist of three folds: the outer one, next to the shell, is involved in shell secretion; the middle one has a sensory function, and the inner one is muscular and controls water flow in the mantle cavity. The mantle is directly attached to the shell by muscle fibres along the pallial line that runs in a semicircle at a short distance from the edge.

Gills. Gills are two large lamellar structures attached to the dorsal margin of the mantle and fill up the mantle cavity to maximize the surface exposed to the inhalant water flow. The large vascularised surface makes gill well suited for gas exchange. Gills are covered of cilia used to convey food particles to labial palps and mouth. (Cannuel R, 2009).

Foot. Ancestor molluscs had a ciliated flat foot covered of mucous glands used as motion organ. After the development of the shell, foot of bivalves lost much of its motion role and became smaller and anterior directed. In some bivalves that have developed a byssal apparatus, such as *M. galloprovincialis*, foot regained a primary biological role. The foot consists of layers of circular and longitudinal muscles, its ventral surface is covered by cilia and several kinds of glands with specific roles in crawling and attachment.

Attachment features. The mollusc byssal apparatus evolved to anchor larval organisms during metamorphosis. Some species of mussels, such as *M. galloprovincialis*, have retained the byssus also in the adult stage. Mussel uses foot to find a suitable substrate, and secretes a single byssal thread. As the mussel grows in length more threads are secreted. Byssus thread is a proteinaceous structure that consists of a root attached to the retractor muscles, a resistant and resilient stem, and a plaque to adhere to surfaces. The main components of byssus threads come from collagen family and foot proteins family and are produced by the foot glands. Proteins are stockpiled in the foot and secrete directly on the forming byssus or released into the byssal groove where a template for thread and plaque is ready for further use. The resistance of the stem are due to the introduction of metal mediated cross-links between the tyrosine residues of foot proteins. Different cross-linking degrees as well as variable gradients of metal ions and collagen types along the stem modulate its physical properties. Also the plaque adhesive

properties are mainly due to metal mediated cross-links; moreover different substrate leads to secretion of alternative variants of foot proteins. (Sun CJ, 2005).

Digestive system. Mussels are filter feeders organisms. Mouth is located at the anterior region below the hinge and has a pair of triangular shaped labial palps. The main function of the labial palps is to remove the inalated material from the cilia of gills to prevent gill saturation (Bennet-Clark HC, 1976). The mouth and the oesophagus are ciliated to guide particles towards the stomach, that is completely surround by the digestive gland and is connected with it through several ducts. Within these ducts there is a two-way flow: materials enter the gland for digestion and absorption by pinocitosis and wastes are enclosed in excretory spheres and swept away through the exhalant opening. The digestive gland is also an important site of storage of metabolic reserves that are used as energy source during the gametogenesis process.

Gonads. Gonad, developed within the mantle tissue, is a system of branching tubules and gametes that are budded off the epithelial lining of these tubules. The tubules convey to a short gonoduct that open into the mantle cavity, then gametes are shed through the exhalant opening. In Mytilus galloprovincialis the mantle filled with gametes is typically orange in females and creamy-white in males.

Heart and haemolymph vessels. The heart is localized in the dorsal region of the body, close to the hinge. It is a single muscular ventricle with two thin auricles. Haemolymph flows from the auricles into the ventricle, that contracts to drive the haemolymph into the aorta. The aorta divides into many arteries that supply mantle, stomach, intestine, shell muscles, and foot. Then, haemolymph is collected into three extensive spaces, the pallial, pedal and median ventral sinuses. Therefore, the circulatory system is also an open system with the haemolymph in the sinuses bathing the tissues directly. From the sinuses the haemolymph is carried to the kidneys for purification, then enters the gills. and comes back to the kidneys to return to the auricles of the heart. Haemolymph has the same main function of blood in mammals, including gas exchange, osmoregulation, nutrient distribution, waste elimination and internal defence. Moreover, it also has a role of fluid skeleton, giving temporary rigidity to the labial palps, foot, and mantle edges. The haemolymph contains cells called haemocytes that have no respiratory pigment because the haemolymph oxygen concentration is similar to that of seawater.

Nerves and sensory receptors. The nervous system is symmetrical and consists of three pairs of ganglia. Visceral ganglion is located on the surface of the posterior adductor muscle and also controls gills, heart, kidney, digestive tract, gonad, and the posterior part of the mantle. Pedal ganglia is localized near the foot and controls it. Finally, cerebral ganglia innervate the palps, anterior adductor muscle, and the mantle. All the ganglia are joined together by pairs of symmetrical nerves (Lon A, 2007). During the evolution, bivalves lost a distinct head, so most of the sense organs are moved to the middle fold along the edge. There are sensory cells sensitive to touch, chemoreceptors, and ocelli that can detect changes in light intensity (Wilkens LA, 1991).

1.1.3 Ecological aspects of *M. galloprovincialis*

Mytilus genus resides in the intertidal and subtidal regions, grouped in dense beds of hundreds of individuals attached to a wide variety of substrates, for example rocks, stones, wood, shells, and also cement or other human-made structures. Several physical and biological factors affect mussel growth and distribution.

Physical factors. M. galloprovincialis optimal growth temperature ranges from 10 to 20°C, but with a high degree of tolerance especially for higher temperatures. The optimal water salinity is about 30‰; the organism is sensitive to higher percentages, but can tolerate lower salinity up to 5‰. For subtidal organisms additional physical factors play an important role such as water depth (up to 4 m), turbidity, and currents.

Biological factors. Diet consists of a variety of suspended particles (named seston) such as bacteria, phytoplankton, micro-zooplankton, detritus, but also dissolved organic material (DOM), such as amino acids and sugars. *M. galloprovincialis* is characterized by a retention efficiency of 90% for particles up to 3 µm diameter, but is also able to catch algae with higher diameters. Various studies have shown that there is both spatial and seasonal variation in the quantity and quality of seston in coastal waters; but, it was estimated that the seston quantity commonly varies from 3 to 100 mg/l, of which 5–80% may be organic. Distribution is also affected by the presence of predators (such as gastropods, crabs, and birds) and pathogens and symbionts that affect the mussel physiology or the structural integrity of the shell.

1.1.4 Reproductive cycle

Mytilus galloprovincialis becomes sexually mature during the second year of life. It undergoes an annual reproductive cycle which involves a period of gametogenesis

followed by two spawning events, followed by a period of gonad reconstitution. Temperature and food availability are the exogenous factor that mainly influences the timing of gametogenesis and spawning: in fact spawning period is highly variable also within the same species in different geographical localizations. In Adriatic sea *Mytilus galloprovincialis* spawns from October to February with two main events in December and February.

Spermatogenesis results in flagellated spermatazoa that measure about 50 µm in length. During oogenesis the oocytes undergo a period of accumulation of lipid globules and small quantities of glycogen called vitellogenesis, that leads oocytes to reach 70 µm in diameter. These reserves are used during the first two larval stages, then, once their feeding apparatus will be functional, larvae are totally dependent on plankton for food. Eggs and sperm are shed directly from the gonoducts into the water column, where fertilisation takes place. In *M. galloprovincialis* optimal conditions for fertilization are a temperature of about 18°C and water salinity ranged from 25 to 35‰.

1.1.5 Larval development

Larval development consists of six stages; the first four are planktonic, followed by settlement and metamorphosis into the adult organism (Figure 1.2). In optimal conditions 25-30 days correspond to larval life. The fertilised egg rapidly divides and cilia start to appear after 4–5 hours. A ciliated trochophore stage of 90 µm is reached about 24 hours after fertilisation. During the second day the first larval shell (named prodissoconch I) begins to form originating from ectoderm; and after 24-36 hours is completed. The second larval stage is named D-larva because its shell is “D” shaped. During this stage mantle, muscles and the first simple feeding apparatus develop. After about eight days, the larva, now called veliger (100–120µm shell length), is provided with a velum, a circular lobe of tissue bearing a ring of cilia that serves as swimming and feeding organ. During this stage begins the secretion of the second calcitic and aragonitic larval shell (named prodissoconch II) and the development of the digestive apparatus, gills, nervous system and byssus gland. Umbo-stage is reached 12 day after fertilization and is characterized by the formation of a rounded knob, called “umbo”, from the hinge. Organs development continues but it is associated with a very fast growth that leads the organism to nearly double its size in only five days (from 150 to 250 µm). After 17-20 days, the fifth larval stage, called pediveliger, is characterized by: a light sensor called eyespot, an extensible ciliated foot used for crawling and byssus secretion; the cerebral, pedal and visceral ganglia, and a complete digestive system

including ciliated palps, oesophagus, stomach, a large digestive gland, and simple intestine. The larva is also provided with few pairs of gill filament but they do not still play their respiratory role. Pediveliger drops from plankton to seabed, responding to light and gravity; then it uses its foot to select a suitable substrate for settlement. This event marks the end of planktonic life and the beginning of metamorphosis. During metamorphosis mussels undergo a very fast spatial organs reorganization to prepare the organism for the adult sessile life. A lot of changes occurs, for example the mantle cavity opens and the water starts to circulate; gills acquire their respiratory role; and begins the secretion of the first adult shell, called dissoconch. This shell is characterized by the peculiar black covering and by the shift of the growth axis responsible for the adult asymmetric shape.

Figure 1.2 : *Mytilus galloprovincialis* larval development. The six larval developmental stage of *M. galloprovincialis.* For each stage, time from fecundatioton, median length, and peculiar features are reported.

1.1.6 Ecological and economical relevance of *Mytilus galloprovincialis*

Biomineralization. Mussels are accumulator of calcium and carbon that in the form of calcium carbonate is the main component of the shell. There are a lot of chemical and biological studies to understand the mechanisms involved in mussel biomineralization process, how biological molecules are able to induce the specific crystallization of calcite or aragonite, and organize the crystals in complex structures (Fang D, 2011; Kinoshita S, 2011). These studies could help us to better understand the molecular mechanisms underlying similar processes of tissue mineralization such as the formation of the bones.

Pollution biomonitoring. Filtering water mussels accumulates a wide range of water contaminant in their tissues. Because of their function, digestive gland and gills are the major sites for the uptake of metal and organic contaminants. Metallothioneins, antioxidant enzymes, and oxyradical scavengers have high activities in these tissues, and are good markers to estimate the degree of pollution. The response of these tissue to different contaminant exposure has been well documented for *M galloprovincialis* (Raftopoulou KE, 2012; Furdek M, 2012).

Immune system. The innate immune system is an ancestral process preserved along the evolution until vertebrate but outclassed by the development of the adaptative immune system. Immune defense in mollusks is guaranteed only by the innate immune system so they are suitable organisms for the study of this biological process (Vera M, 2011; Venier P, 2011).

Economic role. The annual world mollusks production is estimated at 12 million tons, about the 25% of the entire aquaculture production, with a estimated economic value of about 10 billion dollars. In agreement with the FAO (Food and Agriculture Organization of the United Nations, http://www.fao.org/) the annual production of *Mytilus galloprovincialis* is about 1 million tons, mainly from China and Spain. Producers are particularly interested in studies about diseases and parasites that can cause abrupt drop of production efficiency (Francisco CJ, 2010; Longshaw M, 2011); but also in the development of new techniques of aquaculture and restocking of areas damaged by disease, storms or human intervention (Carl C, 2011; Genovese G, 2012).

Adhesive properties. In few seconds, the byssus is able to adhere to a large range of surfaces creating strong chemical bonds in aqueous environment. Glue industry is

particularly interested in studying this natural marine polymer to mimic its surprising properties(Lee BP, 2011). On the other hand, naval industry tries to find new strategies to avoid mussels fouling to keels that slows down the ships and increases fuel consumption (Marcheselli M, 2011).

1.1.7 Genomic, transcriptomic, and proteomic information about *M. galloprovincialis*

In spite of the great interest about morphological, ecological, and economical aspects of *M. galloprovincialis*, little is known about the molecular bases of fundamental bivalves processes such as the regulation of growth and differentiation or sexual maturation by the lack of information about mussel genes and genome.

Genome. Mytilus galloprovincialis genome has not been yet sequenced, as well as the genomes of the other *Mytilus* species. DNA content is estimated in 1.41–1.92 pg (www.genomesize.com) organized in 14 pairs of chromosomes homogeneous in size (Saavedra C; 2006). On the other hand, several copies of mitochondrial genome have been deposited. Mussel mitochondrial genome is studied because shows two different forms, F and M, which are transmitted by female and male individuals respectively. The two mitochondrial sequences of about 16,744 nt diverged by 20%, but preserved identical gene content and arrangement (Obata M, 2011).

Transcriptome. In 2009 Venier *et al.* (Venier P; 2009) have defined the first species specific EST catalogue of *Mytilus galloprovincialis*, including 7112 unique transcribe sequences obtained by the sequencing of 17 tissue specific libraries from adult samples using Sanger technology. All these consensus sequences annotated by Blast search and Inter-ProScan were deposited on Mytibase, a species specific EST database. More recently, Craft *et al.* (Craft JA; 2010) used 454 pyrosequencing to generate 175,547 reads from tissue specific samples, but they have not deposited the resulting contigs. To date, 19,754 *Mytilus galloprovincialis* EST sequences, with a high rate of redundancy, are deposited on NCBI database (www.ncbi.nlm.nih.gov, January 2012), these are the 29% of all the ESTs publicly available for the Mytilus genus.

Proteome. A total of 6,460 proteins are deposited for the Mytilus genus, and about the 25% consists of *Mytilus galloprovincialis* sequences. But these sequences are characterized by a high rate of redundancy.

To date, molecular information directly obtained from larval samples is not available. There are only few studies that demonstrate the larval expression of genes involved in the adult innate immune system.

Several studies have demonstrated that the world wide distribution of the *Mytilus* genus results in a high genetic distance between the species, but also between the different populations of the same species (Gerard K, 2008). This observation explains the high rate of redundancy of sequences on databases, and makes difficult to use sequences obtained from other laboratories around the world; for example to clone a specific gene or to lead an assembly.

1.2 Analysis of the transcriptome

Over the past decade our understanding of transcriptome have radically changed. The first revolutionary discovery was the production of multiple transcript variants from each gene, leading to a dramatically increase in the scale and scope of genome output. The second revolution came with the discovery of non-coding RNAs and their role in cis- and trans-regulation of genes activity. In this context, Sanger methods is unsuitable to represent the complexity of the transcriptome, due to the time consuming physical set up of a cDNA library and the production of few sequences during each sequencing run. The advent of massively parallel next-generation sequencing, which produce millions of smaller sequences, it has allowed a more complete analysis of the transcriptome, identifying the lower expressed transcripts and all the possible variants of a single gene (Eveland A, 2008). To quantify the potential of these new technologies, it was estimate that the 25% of transcripts observed in next-generation RNA-Sequencing experiments is not represented by current gene models (Sultan M, 2008).

1.2.1 Comparison between microarray and RNA sequencing technology

The first genomic platform to analyze expression levels of thousands of genes simultaneously was microarray technology that allows rapid discovery of gene pathways involved in biological processes and pathological states. One of the main advantages of this platform is represented by the comparison of gene expression levels of a large group of gene of interest between a lot of samples. However, microarray technology is characterized by several limitations. The first and more important is represented by the requirement of extensive *a priori* knowledge about the transcriptome

of interest for a successful probes design. Another limitation is the difficult to design specific probes that are able to distinguish between gene variants and do not result in cross-hybridization. The last limit relates to the difficult to define the expression levels of rarely expressed transcripts.

Alternatively to microarray technology, the RNA sequencing (RNA-Seq) has been proposed to define gene expression profiles. Sequences (reads) produced by RNA-Seq are mapped to the reference genome and counted to estimate expression levels of genes. RNA-Seq is not affected by the same limitations described for microarray technology. First it has a greater dynamic range in expression levels because read counts are not affected by the same saturation and sensitivity limitations observed for array fluorescence signals. Furthermore, novel mRNAs, transcripts with alternative start points, alternative polyadenylation sites, and splicing events can be detected (Malone JH, 2011). However, this approach is also characterized by several limitations. First, next-generation sequencing generates great abundance of data that are stored into files of about 20-30 GB and bioinformatics support is fundamental to process and analyze sequencing data. Another limitation is represented by the coverage of sequencing that is requested to correctly represent an entire transcriptome. For highly expressed genes, small number of reads is sufficient; but, for the middle and low expressed transcripts, it is necessary to produce a large quantity of sequences. Next generation sequencing and microarray platforms share two problems: the detection of low expressed genes and the need for *a priori* knowledge about the transcriptome/genome, to design microarray probes and to map small reads. These problems represent technical limits for microarray technology, but they can be overcome with next generation sequencing technology. In fact, to detect rare transcripts is sufficient to increase the number of sequences obtaining higher coverage. Anyway, if genome information is not available the detection of transcript variants of the same gene will be difficult aligning short reads. Furthermore, complex algorithms and approximations have to be performed to estimate the actual level of gene expression because reads counting can not be normalized on gene length. All these problems can be faced only with high power calculators and rigorous statistical approaches (Martin AJ, 2011).

Table 1.1 : Comparison of next-generation sequencing platforms

Generation	Platform Name	Company	Method of Sequencing	Method of Detection	Median Read Length [nt]	Read Number [nt]	Bases *per* Run	Run Time	Cost	Advantages	Limits
First	**3730xl**	Life Technology /ABI	CE-Sanger	Optical/Fluorescence	1,000	94	68,000	1 h	Instrument: 100,000 $; High consumables cost	Long resd length, good ability to call repeats and homopolymers	High cost; Low output.
Second	**Genome Sequencer FLX+ System**	Roche	Pyrosequencing	Optical	850	1 M	700 M	23 h	Instrument: 500,000 $; High consumables cost	Long read length	Difficulties reading homopolymers.
Second	**HiSeq2000**	Illumina	Reversible terminator sequencing by synthesis	Optical/Fluorescence	2 x 150	2 G	600 G	11 days	Instrument: 650,000 $; High consumables cost	Very high throughput	Long run; Significant cost of data analysis.
Second	**5500xl**	Life technology /ABI	Sequencing by ligation	Optical/Fluorescence	50	2.4 G	140 G	7 days	Instrument: 700,000 $; High consumables cost	Very high throughput; high accuracy.	Significant cost of data analysis.
Second	**Heliscope**	Helicos	Single molecule sequencing by reversible terminator synthesis	Optical/Fluorescence	35	800 M	30 G	30 hours	Instrument: 1,000,000 $	Single-molecule nature of sequencing	Short reads increase cost and reduce quality of assembly
Third	**PacBio RS**	Pacific Biosistem	Real time single molecule sequencing by synthesis	Optical/Fluorescence	1,100	150,000	200 M	6 hours	Instrument: 700,000 $ High consumables cost (flow cell)	Single-molecule nature of sequencing; read length up to 3,000 nt	Inefficient loading of DNA polymerase in ZMWs
Third	**Ion Torrent**	Life technology	Sequencing by synthesis	Change in pH detected by FETs	200	5 M	1 G	2 hours	Instrument: 100,000 $;	Direct detection of nucleotide incorporation event	Difficulties reading homopolymers.

Therefore, next-generation sequencing and microarray are two complementary high-throughput approaches to evaluate gene expression signatures with similar performance and opposite limits (Liu S, 2011). For example microarray platforms are more useful to define expression profiling of marker genes or transcripts involved in specific biological process across an high number of biological samples; instead, RNA-Seq provides a complete overview of the transcriptome associated to a model and non model organisms.

Several next generation platforms based on different technologies and characterized by peculiar properties have been developed (Table 1.1). In the following paragraphs we described the most used next-generation sequencing platforms as well as the rising technologies.

1.2.2 Second-generation massively sequencing platforms

All these second generation sequencing platforms are based on the concept of cyclic-array sequencing that can be summarized as the sequencing of a dense array of DNA or cDNA molecules by iterative cycles of enzymatic manipulation and imaging-based data collection.

Roche 454 Genome Sequencer FLX system. This system is based on sequencing-by-synthesis with pyrophosphate chemistry, it was developed by 454 Life Sciences (Roche) and was the first next-generation sequencing platform available on the market. cDNA sample is fragmented and two adapters are ligated to small fragments. Sample is denaturated, diluted, and immobilized on capture beads. The process is optimized to guarantee that a single fragment is bound to each bead. The bead library is emulsified with the amplification reagents in a water-in-oil mixture in a reaction called “emulsion-PCR” (Dressman, 2003). Each bead is captured within its own emulsion micro reactor, where the independent clonal amplification of its bound fragment takes place. For sequencing, beads are layered on a micro-welled plate where each well contains one sample bead. At each cycles sequencing reagents and the nucleotides flowed sequentially in a specific order across the wells of the plate. In pyrosequencing reaction each incorporation event results in light emission detected by a CCD camera. Across multiple cycles, the pattern of detected incorporation events reveals the sequence of the cDNA fragment. A major limitation of the 454 technology relates to resolution of homopolymer, because dNTPs have not terminating groups to prevent multiple incorporations at each cycle, pyrosequencing detects fluorescent signal to determine the

number of repetitive incorporations and this is characterized by a great error rate. As a consequence, the dominant error type for the 454 platform is insertion-deletion (Rothberg and Leamon, 2008). Compared with other platforms 454 sequencer produce few reads with a higher per-base cost of sequencing. The key advantage of the 454 platform is read length. This system can generate about 1,000,000 individual reads with average length of 400 bases. Moreover the new version of the instrument, the Genome Sequencer FLX+ system, provides the same number of reads but characterized by a length up to 1,000 bases, such as Sanger method.

Illumina Solexa sequencing platform. This system is based on sequencing-by-synthesis and now is the most widely used platform by research community. cDNA sample is fragmented and two adapters are ligated to the ends. Small fragments are denaturated, diluted, and immobilized on the surface with flow cell creating a random array. The surface is coated of the adapters sequences that act as primers for the localized clonal amplification of each fragment by bridge PCR (see Figure 1.3)(Adessi C, 2000), resulting in millions spots array where each spot consists of about 1,000 copies of a single fragment. The reaction mixture for the sequencing contains four reversible nucleotide terminators, labeled with four different fluorescent dyes. After incorporation, the nucleotide terminator and array position of each spot are detected and identified with a CCD camera. Then the 3'-end terminator group and the fluorescent dye are removed and the synthesis cycle is repeated. Read-lengths are limited to 150 nt by multiple factors that cause signal decay such as incomplete cleavage of fluorescent dye or terminating group. The dominant error of this technology is substitution with an error rate of 1.5% . However, Illumina platform has two key advantages: the high amount of reads generated and the possibility to repeat the sequencing after a stripping step, where synthesized sequences during the first sequencing run are discarded to allow the restarting of sequencing reaction from the other end. Using the last version of the instrument, the HiSeq2000, a complete sequencing run generates about 2 billion reads of 2x150 nucleotides.

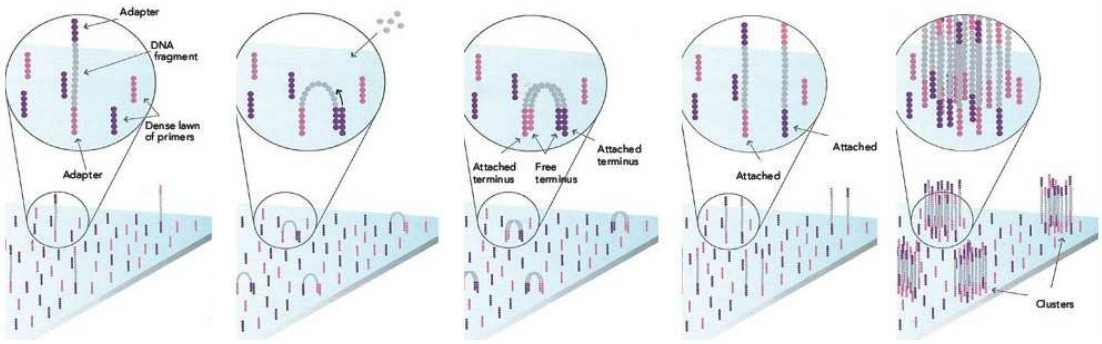

Figure 1.3 : Bridge PCR reaction. Single-stranded fragments randomly binds to the surface of the flow cell covered with primers. Another primer starts PCR. The enzyme incorporates nucleotides to synthesize double-stranded bridges on the solid-phase substrate. Denaturation leaves single-stranded templates anchored to the substrate. After several PCR cycles, millions of dense spots of double-stranded DNA are generated.

Life Technologies SOLiD system: This platform is based on sequencing by ligation technology. Clonal amplification is performed by emulsion PCR, as described for 454 sequencing. Beads bearing amplification products are recovered and immobilized to the surface of the flow cell to generate a dense, disordered array. The sequencing methodology is based on sequential ligation with dye-labeled oligonucleotides (Figure 1.4). First, a universal primer is hybridized to the adapter sequence. Next, a set of four fluorescently labeled oligonucleotide octamers compete for ligation to the sequencing primer. These octamers are degenerated, but the two first bases are correlate with the color of a fluorescent dye at the end of the octamer. So, the fluorescent detection of the first incorporated octamer allows to identify the first and second nucleotides of the sequence. The ligated octamer oligonucleotides are cleaved off after the fifth base, removing the fluorescent dye, then hybridization and ligation cycles are repeated, determining nucleotide 6 and 7. Following a series of ligation cycles, the extension product is removed and the sequencing will restart using another universal primer complementary to the n-1 position for a second round of ligation cycles. After five rounds of sequencing each base is interrogated in two independent ligation reactions.
This method is called ‘Two Base Encoding’ and guarantees a low error rate reducing systemic noise but limits the read length to only 50 nt. The last version of the instrument (SOLiD 4 system) produces about 2.4 billion reads *per* run.

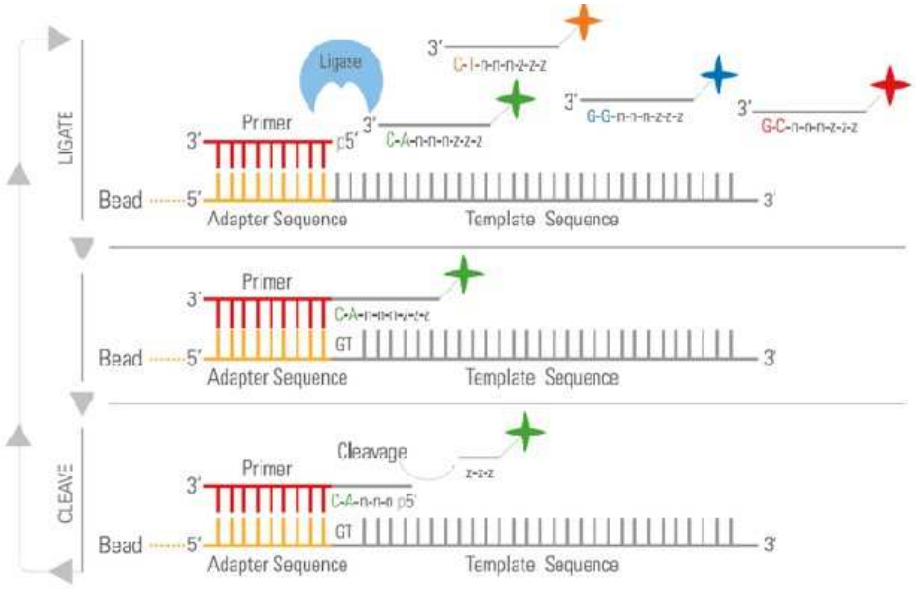

Figure 1.4 : Sequencing by ligation in Solid platform. Primer hybridizes to the adapter sequence. A set of four fluorescently labeled di-base probes compete for ligation to the sequencing primer. Specificity of the di-base probe is achieved by interrogating every 1st and 2nd base in each ligation reaction. Multiple cycles of ligation, detection and cleavage are performed.

1.2.3 Applications of next generation platforms

All these technologies have been used for different applications, according to their technical properties.

To find SNPs and variants by mapping a transcriptome on a reference genome the best approach is to use the small but high reliable reads generated by SOLiD system that allow to rapidly obtain an high coverage (Mckernan KL, 2009); and similar results can be obtained also with Illumina (Mortazavi A 2008). Instead this same approach has not proved successful in the absence of a reference genome and using sequences of related species to lead the assembly (Everett MV, 2011). However, the short reads give a very limited view of the complete transcript, so 454 sequencing is preferred to identify genomic alterations, specifically gene fusions (Grossmann V, 2011).

For trascriptome *de novo* assembly of non model organism 454 sequencing technology is mainly used (Zeng V, 2011) because the longer reads make the assembly step more easy also with a lower coverage. However, the best approach is a hybrid assembly strategy with reads produced by 454 technology and Illumina or SOLiD. Short reads produced in large amount (Illumina or SOLiD) are assembled into contigs, and long reads (454) are subsequently used to scaffold the contigs and resolve variants (Dalloul R, 2010).

Several protocols were set up to quantify gene expression, inspired to the tagging approaches, but they were adapted to new sequencing technologies in order to maximize the number of tags and assay accuracy. Some examples are nanoCAGE (Plessy C, 2010), PEAT (Ni T, 2010), and deepSAGE (Hestand MS, 2010) performed using Illumina or SOLiD (Siebert S, 2011). From a commercial point of view Lyfe Technologies and Illumina are focusing on RNA-Seq. In fact, Lyfe Technologies has recently presented SOLiD SAGE System to determine mRNA expression levels by sequencing unique sequence tags of 27 nucleotides isolated from the 3' ends, whereas Illumina suggests a paired end approach. 454 platform is less used with tagging approaches to avoid the reintroduction of concatamers. But, it is used to define gene expression signatures in non model organisms because the longer reads produced are useful for assembly and annotation processes (Eo SH, 2012; Mahomed W, 2011).
Moreover, Solid and Illumina are the most used platforms to discover and detect expression levels of microRNA; while 454 is used in amplicon sequencing (Rosani U, 2011; De Schrijver JM, 2010).

1.2.4 Single molecule sequencing platforms

In the past years antisense transcription was considered biological or technical noise, now it is clear that many antisense transcripts have functional and various biological roles. Therefore, there is an increasing interest in determination of antisense expression profiling. Standard RNA-seq approaches generally require double-stranded cDNA synthesis, loosing RNA strand information. So, several protocols have been developed to select only the sense cDNA strand (Cloonan N, 2008), or performing the ligation of adaptors in a predetermined orientation to the ends (Lipson D, 2009). Levin et al (Levin JZ, 2010) compared the principal strand-specific RNA sequencing methods revealing several substantial difference in obtained results. These differences are mainly due to biases and artifacts specifically introduced by ligase and retro-transcriptase and then amplified by PCR. It has been demonstrated that ligation reaction showed sequence preferences (Faulhammer D, 2000), and unexpected second-strand cDNA artefacts can be introduced during first-strand cDNA synthesis due to DNA-dependent DNA polymerase (DDDP) activities of reverse transcriptases (Perocchi F, 2007). So the only method to avoid these problems is represented by a single molecule sequencing of cDNA without the need for the amplification step. Moreover, single molecule sequencing, without the amplification step that could affect the relative representation of

transcripts, could be a good choice for the study of gene expression. At the present only two available single molecule sequencing platform are reported.

HeliScope Genetic Analysis System. is the first single molecule sequencing system and recently available on market. Single-stranded cDNA library is disorderly arrayed on a flat substrate without any amplification. DNA polymerase incorporates one of four fluorescently labeled nucleotides at each sequencing cycle and a CCD camera detects the signal. After washing, fluorescent dye is chemically removed, and another cycle of sequencing could be performed. The major limitation of this technology is the sequencing accuracy due to difficulty to detect a single molecule event. Therefore, the main error is deletion but a two-pass strategy can reduce this error rate. Anyway, this problem as well as the incomplete cleavage of fluorescent dye or terminating group limit reads length to only 35 nucleotides. For these reasons as well as the high costs (1 million dollars) this tecnique is not widely used, but, recently published protocols, that perform single-molecule sequencing directly from RNA without cDNA synthesis (Ozolak F, 2011), probably will increase the number of users.

Pacific Biosciences platform. Pacific Bioscience tried to solve the problem of detecting single molecule event developing nano-structure chip, the Zero Mode Waveguide (ZMW), for real-time observation of DNA polymerization (Levene et al., 2003). ZMW chip consists of thousands of wells working as high parallel confocal imaging systems. Only 150,000 reads are produced due to the complexity of the nano-structure and the stochastic nature of immobilizing a single DNA polymerase at the bottom of each ZMW well. Another limitation is represented by the low accuracy of a single synthesis reaction that can not be repeated. Anyway, this technology will be improved in the next future since it produces 1,000 nucleotides long reads.

1.2.5 Non-optic sequencing

The previous sequencing platforms are based on the detection of a light signal by a CCD camera. Since CCD optic system is the main cost of a sequencing platform, the logical advancement has been to substitute the use of optics for less expensive approach to detection.

Ion Torrent (Life Technologies). This platform is based on sequencing-by-synthesis and uses field-effect transistors (FETs) to measure a pH changes in a microwell structure. The workflow of this technology is similar to 454 sequencing. It is based on an

emulsion-PCR of fragments bound to beads. But the sequencing reaction occurs on a dense array of microwells in which each well is an individual FET detector. In pyrosequencing, when a incorporation event occurs, a light signal is produced, but the reaction also results in a change of local pH and temperature. FETs are able to convert the pH change in a recordable voltage change. Since the voltage scales with the number of incorporated nucleotides, Ion Torrent is able to count repeats with higher precision than 454 sequencing. The last version of the instrument uses the Ion 318 sequencing chip that, in about 2 hours, can produce 5 million reads with an average length of 200 nucleotides. The read length is similar to Illumina platform, instead the number of generated sequences are significantly lower, but the reaction time as well as the reasonably low price (100,000 $) make it the first benchtop-scale, high-throughput sequencing platform and guarantee a future wide distribution.

2 Aims

In spite of the biological and economical relevance of marine bivalves, developmental biology research is limited by the poor availability of public resources such as deep-coverage EST database or sequenced genomes. We want to increase the knowledge about this crucial, but often overlooked, stage in the life of mussels focusing our attention on the dissection of complex mechanisms that characterized its progression. We want to identify key genes involved in the fundamental biological processes of the organism, such as shell synthesis or byssus secretion. This work want to expand transcriptional knowledge of *M. galloprovincialis*, focusing on larval expressed transcripts, and characterizing their expression profiles across seven principal larval stages.

3 Materials and Methods

3.1 Primers table

Primers for libraries construction and for 5’ and 3’-RACE

Primer name	**Sequence 5’→3’**
Oligo(dT)15-BpmI-T7	CACACACATAATACGACTCACTATAGGCTGGAGT_{15}V
SMART-T3	CACACACAATTAACCCTCACACTAAAggg
biotinilated-T7	biotin-CATAATACGACTCACTATAGG
T3	AATTAACCCTCACTAAAGGG
Oligo(dT)-IIA	CAAAGCAGTGGTATCAACGCAGAGTACT_{20} VN
Primer II A	AAGCAGTGGTATCAACGCAGAGT
T3Long	CACACAATTAACCCTCACACTAAAGGG
the lowercase letters represent ribonucleotides ; “V” and “N” are IUPAC nucleotide codes for degenerated bases: V= A or C or G; N= A or C or G or T.	

Primers for qRT-PCR

Annotation	**Isogroup ID**	**Forward primer 5’→3’**	**Reverse primer 5’→3’**
Actin	00268	AAAGCTGGATTTGCAGGAGA	GGGTCTGCCAACAATGGAT
Calcium binding protein	05285	GCGGTCTCATTGTATGGCTTA	TTCCTAGTTGTTTGTTGTCTCTTAGG
c-myc	00180	GGTCAGAAATGGTCAGATGGA	ACATACGTTTGGTAAATTCCATGTA
Contactin	00517	TAACTGCACATGTGGCCAAG	AAGGCGTTAGTCGGACAATTT
Dynein light chain	00033	GAATGGAAAATGAAATGTGAACAG	TGTGTGTTTCTTTTCTTCCACTTG
Elongation factor 1	03408	TTGAAACCAACATTGTCTCCTG	TCCGTAGAAATGCACCATGA
MSP-130	02692	CGTATACCCGCTGGAATGTT	AATATCCTCACTCTTTTGGTGGTC
Shematrin	02203	CCTCCAACACCACTGTAGCTT	GCAACTATCTATACGGAAATTCTGG
Sym 32	00054	GGTTCTAAAATCTCCAAATTTGACC	GAGGCATCATGACCTTTTCAA
Tubulin α	00008	AAATCTTCACGGGCTTCTGA	CCATTGGTACGTCGGAGAAG
Unknown	00514	TTTTACCAAACCGCACACAC	CACCAACTGTTCCTCTTGCAG
Vdg3	00306	CATTCGGCTTCACATGCTT	AGATGGGATTGTAGGAACATTAGG
18S rRNA	we used the Quantum RNA 18S Internal Standards (Ambion)		

Primers for *in situ* hybridization

Annotation	**Isogroup ID**	**Forward primer 5’→3’**	**Reverse primer 5’→3’**
Shematrin	02203	ACCGGCATAATATCCACCAC	GCGGATATGGCTATGGAACT
Vdg3	00306	TCGTATTTAAACAGTTGGTGCAA	CAACTTCAAGATGAACTTGCTTT

3.2 Mating experiments

3.2.1 Facilities

We have performed spawning and larval rearing at the CRIM (*Centro Ricerca Molluschi*) laboratory of the "*Istituto Delta"* in Goro (Fe) that is a spin-off company of the university of Ferrara (http://www.istitutodelta.it/). This centre, specialized in the induction of bivalve mating, is equipped with an hatchery room, an algal culture facility, and a seawater filtering system.

Hatchery room - It is a room kept at the optimum temperature for mating (16-18°C). It is equipped with non-toxic polyvinylchloride tanks, different size filters, and all the equipments needed to handle large amounts of water.

Algal culture room - A batch culture system is used for algal production to ensure the availability of fresh food during the entire duration of the experiment. Required species were inoculated in 25 liters spherical glass flasks filled with filtered enriched seawater (Guillard's Marine Water Enrichment Solution, Sigma) at 20°C, and it is exposed to a continuous artificial light source. Each flask is aerated with a air/carbon dioxide mix; 2% CO_2 assures the carbon source for photosynthesis and maintain the pH within the optimal range (from 7.5 to 8.2). Under these conditions the culture grows quickly until light permeation is inhibited by the high cell density. Then an aliquot is inoculated in a new flask to continue production.

Seawater filtering system - The intake of the seawater system is placed several meters below the sea surface to avoid any floating contaminants such as petrochemicals and plastics. Incoming water is first passed through sand filters that eiminate most particulate material greater than 20-40 µm in size. Then the water is filtered with 0.2 µm filter to remove organisms, bacteria and algae that could be detrimental to larval life. Salinity is measured using a hydrometer and adjusted to 29.5%. The optimal temperature of the water is reached by incubating the vessels filled with water in the hatchery room for at least one day.

3.2.2 Spawning

Mating experiment was carried out at February 2009 and repeated at February 2011 according with the protocol set up by Dr Edoardo Turolla (CRIM laboratory, Goro). About 80 adult mature *Mitilus galloprovincialis* of 5-6 cm of length were taken from a mussel farm located 1.5 miles off Goro's coast. Mussels were transported to the CRIM laboratory and cleaned externally to remove any adhering debris. Then bivalves were

conditioned in a filtered seawater flow-through system at the temperature of 16°C until spawning induction. Mussels were placed in the spawning tank, a low and wide tough with a black bottom to provide a dark background to better visualize gametes. Spawning was induced by water oxidative power variation. The tank was filled with filtered seawater to a depth of about 10 cm. To induce the spawning, H_2O_2 was added to a final concentration of 0.5 mM and after 15 minutes water was drained and replaced with new filtered seawater. Generally mussels start spawning within two hours; after 4 hours the animals that did not respond to stimuli were discarded. When a female began to spawn she was transferred to an individual spawning vessel with about 1 liter of filtered seawater. Instead spawning males were placed together in a 5 liters tank. At the end of the spawning, mussels were removed from the tanks. Some eggs from each female was observed with a microscope. Eggs, that showed not of a uniformly dense and granular appearance, or that were not characterize by a well rounded shape, were discarded. Eggs in good conditions, were resuspended and about 5 ml of the sperm suspension were added. After 1 hour some fertilized eggs from each female was observed with a microscope to verify the presence of a single polar body and not multiple fecundation. Good quality embryos were pooled and transferred to a 400 liter tank filled with filtered seawater at a maximum concentration of 80,000 embryos *per* liter. Rearing density has been maintained high during the first days, but it has been gradually decreased during larval development until 5,000 larvae *per* liter for pediveliger stage.

3.2.3 Rearing

Areation - Optimal dissolved oxygen saturation and a proper mixing of the tank water has been guaranteed by an aeration system. A very slow fine bubbling of about 30 l/h was used for the first sensitive stages (embryo and trocophora) to preserve their integrity. Then the aeration rate was gradually increased up to 200 l/h for the following stages.

Feeding - Algal species fed vary with larval size. Embryos and trocophoras do not need feeding. We started to feed larvae at the d-larva stage with Isochrysis galbana which is microalga and easily assimilated. When larvae have reached 150 µm of length we started to use a mix of three algae: Isochrysis galbana, Chaetoceros gracilis, and Tetraselmis suecica. Larvae have been fed after the washing procedure adding algae to a final concentration of 10-15 cells/µl.

Washing – The water was changed once every two days to eliminate waste products during larvae's feeding, respiration, and excretion. Before the water change, larvae were

observed with a microscope to verify their stage of development and their state of health (healthy larvae have brown- yellow coloration with a dark digestive gland). Aeration was stopped to allow debris precipitation to the bottom of the vessel. Water was gently siphoned out the tank into a filter with a mesh characterized by pore size calibrated on the larval size. The larvae were put in a new vessel rinsing the filter.

3.2.4 Sampling

We have collected seven larval stages: 1h embryos, trocophora, D-larva, veliger, umbo-stage, pediveliger, and metamorphic larva. Since larval growth rate depends on many factors such as temperature, nourishment, and water features, it is difficult to define the timing of various stages *a priori*. For this reason each specific stage was microscopically identified by length and the presence of stage specific morphological markers (Table 3.1). Larvae were collected at least 1 day from the feeding to avoid algal contaminations of the sample. Metamorphic larvae and juvenile mussels were detached from the walls of the tank using a soft water jet. Sampling was performed using a filter with pore size suitable for concentrating larvae and discard debris, algae, and any previous larval stages. Larvae were gently rinsed with a soft filtered water jet and three sampling of 100 µl of the dense suspension were collected for each stage.

stage	length (µm)	morphological markers	time from fertilization
embryos	90.0	polar body	1h
trocophora	90.0	ciliated larva	1 day
D-larva	100.0	first shell completed	3 days
veliger	140.0	no unique markers	8 days
umbo-stage	150.0	rounded knob	12 days
pediveliger	250.0	functioning foot and eye spot	16 days
metamorphic larva	300.0	adhesion	22 days

Table 3.1: Specific markers used to select each larval stage. For each developmantal stage are reported the typical length, peculiar morphological markers and the expected time from fertilization under optimal rearing conditions.

The first aliquot was stored at -20°C with 4 ml of TRIzol reagent (Invitrogen) for the total RNA extraction; the second sample was placed at 4°C in 2 ml of Bouin solution for morphological characterization; finally, the third aliquot was stored at 4°C in 4%

paraformaldehyde (PFA) in phosphate buffered saline (PBS) Ph 7.2 (Gibco) for *in situ* hybridization. 250 µl of each larval stage was frozen in liquid nitrogen and stored at -80°C for protein extraction and future western blotting analysis.

3.3 Sample processing

3.3.1 Total RNA extraction and quality control

Samples (100 µl of larvae in 4 ml of TRIzol) were thawed on ice, and homogenated using the ultra-turrax-T8.01 blender (IKA-Werke). 0.2 ml of chloroform were added *per* 1 ml of TRIzol, tubes were shaken by hand for at least 20 seconds, and then they were incubated on ice for 15 minutes. Centrifugation at 12,000 x g for 20 minutes at 4°C separates total RNA in the aqueous phase from the organic phenol-chloroform phase. RNA was precipitated from the aqueous phase by adding 0.5 ml of isopropyl alcohol *per* 1 ml of TRIzol reagent used for the first homogenization. Then samples were incubated at -20°C for at least 30 minutes and centrifuged at 12,000 x g for 25 minutes at 4°C. Total RNA pellet was washed with 500 µl of 75% ethanol, centrifuged 12,000 x g for 15 minutes at 4°C, and dried at room temperature (RT) for 10 minutes. To remove the high amount of carbohydrates, which characterizes mussel samples, pellets were resuspended in 150 µl of nuclease-free water (Gibco) and total RNA extraction was repeated adding 35.5 µl of 8M litium chloride and the solution was incubated at 4°C overnight. After a centrifugation step at 12,000 x g for 25 minutes at 4°C, RNA was washed with 500 µl of 75% ethanol, centrifuged 12,000 x g for 15 minutes at 4°C, and dried at room temperature (RT) for 10 minutes. The total RNA pellet ws resuspended with nuclease-free water to a final concentration of about 1 µg/µl and stored at -80°C.

Total RNA was quantified using NanoDrop 1000 spectrophotometer (Thermo Scientific). 200 ng of each total RNA sample were analyzed using RNA 6000 Nano LabChip on a 2100 Bioanalyzer (Agilent) to estimate RNA integrity. The instrument performs a capillary electrophoresis and shows results as a electropherogram. All poor quality RNA samples were discarded.

3.3.2 Bouin fixation

Samples (100 µl of larvae in 2 ml of Bouin solution) were fixed overnight at 4°C on a rotating wheel. Then larvae were gradually dehydrated with 5 minutes washes with an increasing concentration of ethanol (EtOH/PBS: 30/70, 50/50, 70/30). Samples were stored at 4°C in 500 ml of 70% ethanol.

3.3.3 PFA fixation

Samples (100 µl of larvae in 2 ml of 4% PFA in PBS) were fixed overnight at 4°C on a rotating wheel. Then larvae were gradually dehydrated with 5 minutes washes with an increasing concentration of methanol (MeOH/PBS: 25/75, 50/50, 75/25). Samples were stored at -20°C in 500 ml of 100% methanol until further use.

3.4 Construction of the 3'-end cDNA libraries

Seven independent stage specific 3'-end cDNA libraries were constructed in order to investigate larval development in mussels.

3.4.1 The first-strand synthesis of cDNA using SMART technology

I performed SMART (Switching Mechanism at 5' End of RNA Template) reaction using reagents supplied by the SuperScript II Reverse Tanscriptase kit (Invitogen). 1.2 µg of total RNA was mixed with the Oligo$(dT)_{15}$-BpmI-T7 primer and the SMART-T3 primer at final concentrations of 1 µM and 4,8 µM respectively. Nuclease-free water was added to a final volume of 5 µl, then the mixture was incubated at 72°C for 2 minutes to eliminate any RNA secondary structures. First strand reaction mixture was completed with 1x First Strand Buffer, 10 mM DTT, 0.5 mM dNTPs, 100 Units of Superscript II, and nuclease-free water to a final volume of 10 µl. SMART reaction was incubated for 2 hour at 42°C with the addition of other 100 Units of Superscript II after the first hour of incubation. Finally, reaction was diluted 1:3 and the enzyme was inactivated at 72°C for 7 minutes. Single strand cDNA could be used immediately for the amplification step or stored at -80°C.

3.4.2 cDNA amplification by PCR

Amplification step was performed using Advantage 2 PCR Polymerase kit (Clontech). 1 µl of diluted first strand cDNA was amplified with a T3 forward primer [1 µM] and a 5'-biotinilated-T7 reverse primer [1 µM] in a 50 µl reaction volume according to the manufacturer's protocol. The thermal cycling condition were as follows: 2 minutes of denaturation at 95°C, followed by 22 cycles of 15 seconds denaturation at 95°C, 20 seconds of annealing at 52°C, 2 minute of elongation at 72°C, and a final 5 minutes elongation at 72°C. The number of amplification cycles was defined empirically to produce an adequate amount of cDNA without reaching the plateau of reaction. The resulting cDNA was preliminary assessed by gel electrophoresis (1% agarose). Biotinilated double stranded cDNA was precipitated with 1/10 volume of 3 M sodium

acetate pH 5.2 and 2 volumes of absolute ethanol, and resuspended in nuclease-free water to a final concentration of 1 µg/µl. 400 ng of cDNA was analyzed using DNA 1000 LabChip on a 2100 Bioanalyzer (Agilent) to define the size distribution. cDNA could be used immediately for the amplification step or stored at -20°C.

3.4.3 Fragmentation of double-strand cDNA by nebulization

We performed the fragmentation of double-strand cDNA according to the standard shotgun library protocol suggested by Roche for 454 sequencing. TE buffer (10 mM Tris HCl pH 8.0, 1 mM EDTA) was added to 10 µg of biotinilated cDNA to obtain a 100 µl final volume. 500 µl of Nebulization Buffer (53% glycerol, 40 mM Tris HCl pH 7.5, 5 mM EDTA) were added to the diluted cDNA and mixed by pipetting. Mixture was transferred into a Nebulizer (supplied with the GS DNA Library Preparation kit, Roche) and 30 psi (*per* square inch) of nitrogen was applied through the nebulization chamber for 1 minute. Fragmented cDNA was precipited with 1/10 volume of 3 M sodium acetate pH 5.2 and 2 volumes of absolute ethanol, and resuspended in nuclease-free water to a final concentration of about [1 µg/µl]. We recovered about the 50% of the starting amount of cDNA. 400 ng of cDNA was analyzed with DNA 1000 LabChip on a 2100 Bioanalyzer (Agilent) to verify the library size distribution after fragmentation (300-800 nucleotides).

3.4.4 The 3'-end selection and poly(A) processing

About 5 µg of fragmented biotinilated cDNA was mixed with $5 \cdot 10^7$ magnetic beads (70 µl) conjugated with strepavidin (Dynabeads M-280 Streptavidin, Invitrogen) in 500 µl of Binding and Wash buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl). Biotinylated 3'-end fragments were bound to the beads during 15 minutes of incubation at room temperature on a rotating whell. The 3'-end fragments bound to magnetic beads were collected with a magnetic rack and washed several times with Binding and Wash buffer to remove aspecific fragments. Beads were resuspended in 50 µl of 1x NEB buffer 3 with 100 ng/µl of bovine serum albumin (BSA) and 5 Units of BpmI restriction endonuclease enzyme (NEB) to recover 3'-end fragments and eliminate poly(A). Reaction mixture was incubated at 37°C for 1 hour, then beads were separated with a magnetic rack and discarded. 3'-end fragment were precipitated with 1/10 volume of 3 M sodium acetate pH 5.2 and 2 volumes of absolute ethanol, and resuspended in 10 µL of nuclease-free water. 0.5 µl of resulting 3'-end cDNA libraries were diluted 1:4, and quantified using the Qubit 1.0 Fluorometer (Life Technologies) according to kit

protocol. 3'-end cDNA ibraries were stored at -20°C.

3.5 3'-end cDNA libraries sequencing

Sequencing of the seven stage specific libraries was performed using 454 GS FLX Titanium technology available at BMR Genomics Sequencing Service (Padova). For each library we used an eighth of plate. In brief, cDNA fragments with 5´or 3´overhangs are converted to 5´-phosphorylated blunt-ended fragments, and two adaptors, called Adaptor A and Adaptor B, are ligated to the ends. One adaptor contains a biotin tag that allows binding of the cDNA fragments to beads conjugated with streptavidin. A proper dilution was optimized to guarantee the bounding of only one sequence to each bead. Beads are emulsified with the amplification mixture in a water-in-oil mixture. Each bead is captured within one drop of PCR mixture allowing the clonal amplification of a single sequence. PCR is performed with a biotinilated primer to maintain amplified fragments bound to their specific beads after breaking the emulsion. Beads are mixed with smaller enzyme beads (containing DNA polymerase, sulfurylase, and luciferase) and centrifuged on a plate covered by millions of micro-wells that are able to load only one sequence bead. This plate is placed into the 454 sequencer. Sequencing is based on pyrosequencing technology. A single nucleotide species is delivered across the plate every cycle, where the incorporation of dNTP occurs, pyrophosphate is released and converted to ATP by sulfurylase. Luciferase uses ATP to convert luciferin to oxiluciferin producing light. A CCD camera captures the image of the plate after each cycle and the GsRunBrowser (Roche) analyzes the entire succession of images to obtain the final DNA sequence - called "read" - for each well. Using a complex algorithm, the program assigns a score, based on the quality of the pyrosequencing signal, to each read and corrects bad quality reads. Finally the software trims read ends for primer sequences, and returning only the high quality reads.

3.6 Data analysis

3.6.1 Alignments between 454 reads and EST sequences deposited on MytiBase

Alignment between 454 reads and Mytibase sequences were performed using LASTZ (http://www.bx.psu.edu/miller_lab/) at a minimum identitiy of 95% and at least 90% of coverage.

3.6.2 Assembling of 454 reads

Assembly was performed using Newbler 2.5.1 with default settings for trascriptome de novo assembly. Software makes a preliminary analysis and discards repeats and too short reads (< 20 nt). The first step of assembly is finding overlap between reads. When two different reads have a minimum overlap of 40 bp with at least 90% of identity. When reads overlap, they are used to generate a consensus sequence called contig. Unique sequences, which do not align with any other sequence, are discarded as singleton. If for a given gene there is only one transcript, we obtain a single contig. However, if there are different splicing alternatives, contig behaves as exons (Figure 3.1). For this reason the second step of assembly is creating a contig graph. The software uses reads shared by several contigs to bring them together in a graph that shows all the possible exons combinations. Each of these graphs is called isogroup, and represents a putative gene. Finally, newbler traverses the contigs in the graphs of each isogroup to generate transcript variants, which are called isotigs. Newbler also has a function that allows to align the reads on reference sequences. We used this tool to remap singletons with a reduced minimum overlap (20 bp) on the sequences just assembled in order to reduce the number of unique sequences.

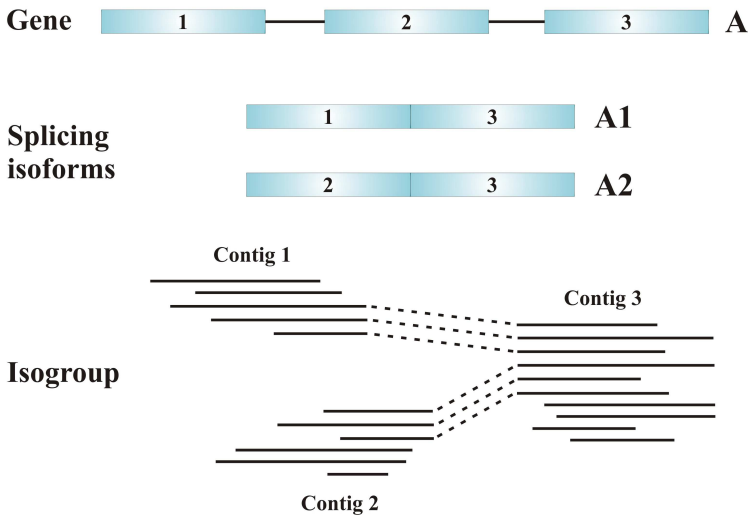

Figure 3.1: Relationship between genes, splicing isoforms, and isogroups. The gene "A" has 3 exons (1, 2, and 3) and it is spliced in 2 different isoforms (A1, A2). Splicing variants will result in reads that share the 3' ends (exon 3) but differ for the rest of the sequence. For this reason each read gives origin to two different contigs. All contigs are grouped together by shared reads resulting in a contig graph (isogroup) that summarizes all isoforms of gene "A". So contigs represent exons. The contig graph is called isogroup and describes gene's structure. Splicing variants are represented by isotig that are the different paths through the contig graph.

3.6.3 Annotation process

We developed an algorithm for the automatic annotation of the isotigs and the singletons generated by Newbler 2.5.1. Each sequence , converted in FASTA format, was searched locally in nucleotide database, dounloaded from NCBI (http://www.ncbi.nlm.nih.gov/), and UniProtUK database (http://www.uniprot.org/), using Blast-X and Blast-N, respectively. First 10 HSPs (High Scoring Pair) from each blast result were collected and stored in a local PostgreSQL table, as a collection of automatic annotation. Each annotation was automatically examined to assign the best description text to the corresponding sequence: matches with expectation values greater than e^{-6} for protein (Blast-X) and e^{-50} for nucleotide (Blast-N) were considered as poorly informative. The algorithm gives priority to the hits and discards the alignment characterized by less than 30% of coverage. Finally, among the five best results, the software selects the description related to the the closer taxonomic distance respect to *Mytilus galloprovincialis.*

3.6.4 Definition of relative abundance of transcripts

We evaluated the gene expression profile of each larval stage by counting the number of reads that align to each isogroup (at least 95% of identity and 90% of coverage) in the six stage-specific libraries. Since libraries have different abundance, in agreement with a recently paper by Bullard *et al*. (Bullard JH, 2010) we have performed a third quartile normalization using edgeR (Robinson MD, 2010).

3.6.5 Clustering analysis

Hierarchical cluster analyses were performed by tMeV, v4.5.1 (MultiExperiment Viewer; Saeed AI, 2006) that is a genomic data analysis tool, incorporating algorithms for clustering, visualization, and statistical analysis. Clustering tree was obtained using the Euclidean distance.

3.6.6 Identification of differentially expressed genes

A negative binomial test was performed to identify significantly differentially expressed transcripts between at least two stages. A binomial test makes a comparison between two categories. We performed all possible combinations comparing single stages or groups of similar stages or single stages *versus* groups. We focused our attention on the comparisons suggested by the cluster analysis. We assumed that an isogroup is differentially expressed in a comparison when the p-value of the negative binomial test

is less than 10^{-3}. The annotation of the differentially expressed genes was manually verified.

3.6.7 Functional annotation

Gene Ontology (GO) analysis was performed on the longer isotig of each differentially expressed isogroup using Blast2GO (Conesa A *et al*., 2005). The program performs a Blast-X analysis against non redundant protein database (nr/nt, NCBI) with an e-value cut-off of e^{-6}. It collects all the GO terms associated to blast results and links functional terms to query sequences. Blast2GO assign to each GO annotation an e-value taking into consideration the similarity between query and blast results, the quality of the source of GO assignments, and other parameters. At the end of the process, we set up an e-value cut-off of e^{-6}. We analyzed third level GO term.

3.7 Quantitaltive Real Time PCR (qRT-PCR)

1.5 µg of total RNA from each stage and pool sample was retro-transcribed with SuperScript VILO cDNA synthesis kit (Invitrogen) according to the manufacturer's protocol. This commercial kit performs the synthesis of the first strand starting from random primers. Single strand cDNA was precipitated adding 1/10 volume of 3 M sodium acetate pH 5.2 and 2 volumes of absolute ethanol, and resuspended in 30 µL of nuclease-free water. 1 µl of retro-transcribed sample was PCR amplified in a 10 µl reaction using GoTaq qPCR Master Mix (Promega) a commercial kit based on the SYBER Green chemistry. Each PCR was performed in triplicate in a 7500 Real-Time PCR System (Applied Biosystem). The thermal cycling condition were as follows: 15 minutes of denaturation at 95°C, followed by 40 cycles of 30 seconds denaturation at 95°C, 2 minutes of annealing and elongation at 60°C, and a final 3 minutes elongation at 72°C. A couple of specific primers for each isogroup of interest was designed using the on-line tool Universal Probe Library Assay Design Center (Roche, https://www.roche-applied-science.com/sis/rtpcr). Gel electrophoresis and the dissociation curve were used to assess the specificity of amplicons. To test the efficiency of each couple of primers, a standard curve was performed starting from 4 serial dilutions of the retro-transcribed pool sample (approximately 10 ng, 1 ng, 0.1 ng, 0.01 ng). Efficiency is equal to $10^{(-1/x)}$ where x is the slope of the linear regression of a dilution *versus* Ct graph (dilution in logarithm scale). Primers characterized by less than 90% of efficiency were discarded and a new couple of primers was designed for the

same isogroup. To evaluate differences in gene expression, a relative quantification method was chosen, where the expression of the target is standardize by a non-regulated reference gene. Actin was used as endogenous control because its mRNA expression level remains essentially constant from sample to sample. To calculate the relative expression ratio (RQ, relative quantification), the $2^{-\Delta\Delta Ct}$ method (Livak KJ, 2001) implemented in the 7500 Real Time PCR System software was used. This method determines the change in expression of a nucleic acid sequence (target) in a test sample relative to the same sequence in a calibrator sample.

3.8 *In situ* hybridization

3.8.1 *In situ* probe design

For each isogroup of interest we designed a couple of primers spaced about 300 nucleotides and with a melting temperature of about 60°C, using the on-line tool Primer3 (http://frodo.wi.mit.edu/primer3/). A high fidelity PCR (Advantage 2 Polymerase Kit, Clontech) was performed according to the manufacturer's protocol starting from a retro-transcribed RNA pool. The thermal cycling condition were as follows: 2 minutes of denaturation at 95°C, followed by 35 cycles of 15 seconds denaturation at 95°C, 20 seconds of annealing at 60°C, 1 minute of elongation at 72°C, and a final 3 minutes elongation at 72°C. Amplified sequence was cloned using T4 DNA ligase (NEB) and pCR 2.1 vector (Invitrogen), which has a T7 promoter near the polylinker. Electro-competent *E.coli* (DH10b) were transformed and several colonies were screened using PCR and Sanger sequencing (BMR Genomics). This cloning strategy is not directional so I selected a colony for both directions in which the insert could be cloned. DNA plasmid were purified with PureLink Quick Plasmid Miniprep Kit (Invitrogen) and 1.5 µg were linearized using SpeI (NEB). *In vitro* transcription was performed with mMESSAGE mMACHINE T7 (Ambion) according to the manufacturer's protocol using the DIG RNA labeling mix (Roche) characterized by a small concentration of digoxigenin-11-UTP. After purification with RNA Clean & Concentrator-25 (Zymogen), labeled RNAs were quantify with NanoDrop, and their quality was verified using 2100 Bioanalyzer as previously described. Resulting antisense labeled transcripts were used as specific *in situ* hybridization probes; while, sense labeled transcripts were used as control sequences. RNA probes and controls were stored at -80°C.

3.8.2 Whole mount in situ hybridization

500 µl were used for all different washes. About 50 larvae from PFA fixed samples were gradually rehydrated with 5 minutes washes in MeOH/PBS-T (1% Tween20 in PBS) with an increasing concentration of PBS-T (75/25, 50/50, 25/75), then larvae were resuspended in PBS-T and washed several times. Pediveliger and metamorphic larvae were decalcified by incubation in Morse solution (10% sodium citrate, 20% formic fcid) for 1 hour, then they were washed in PBS-T several times. Permeabilization was carried out with proteinase K (Sigma) 50 ng/mL in PDS-T at room temperature for a time dependending on larval size: embryos no treatment; trocophora 1 minute; D-larva 3 minutes; veliger 10 min.; umbo-stage 15 min.; pediveliger 20 min.; and metamorphic larva 25 minutes. Larvae were refixed in 4% PFA in PBS for 20 minutes at room temperature to inactivate proteinase K. After several PBS-T washes, samples were prehybridize at 70°C for 5 hours in the hybridation buffer (50% formamide, 5x SSC, 10 µg/mL heparin, 100 µg/mL t-RNA, 1% Tween20, and citric acid 1M to a final pH of 6). Hybridization was performed overnight at 70°C in the hybridization buffer (HB) with 200 µg/mL antisense or sense ribonucleotide probes. To increase the specificity of probes we performed 15 minutes washes at 70°C in hybridization buffer without t-RNA and heparin (wash hybridization buffer, WHB) with an increasing concentration of 2x SSC (WHB/2xSSC: 75/25, 50/50, 25/75, 0/100). After two washes of 30 minutes in 0.2x SSC at room temperature, samples were gradually resuspended in PBS-T with some 10 minutes washes in 0.2xSSC/PBS-T (75/25, 50/50, 25/75, 0/100). Larvae were preincubated with the antibody buffer (2% sheep serum, 2mg/mL BSA in PBS-T) for 2 hours at room temperature, then they were incubated, under gently agitation, with anti-digoxigenin antibody coupled to alkaline phosphatase (Roche) 1:5000 in the antibody buffer at 4°C overnight. After antibody incubation, we performed extensively washes in PBS-T at room temperature. Staining was performed at room temperature with NBT/BCIP (Roche) 20 µl/ml in staining buffer (100 mM Tris HCl pH9.5, 100 mM NaCl, 0.1% Tween 20), and it was monitored using a stereomicroscope. Staining reaction was blocked by resuspending samples in stop solution (1 mM EDTA in PBS pH 5.5). Samples were washed in PBS-T, and stored in 4% PFA in PBS at 4°C in the dark until mounting.

3.8.3 Mounting

Samples were treated with several 5 minutes washes in PBS/Glycerol with an increasing glycerol concentration (80/20, 50/50, 20/80). Using a cutter we obtained a small chamber of about 0.5x0.5 cm on a piece of adhesive tape stuck on a microscope slide. About 10 µl of larvae in 80% glycerol were deposited in the chamber and covered with a coverslip. The thickness of the adhesive tape prevents the coverslip crush larvae. Observations were made with a Leica MZ stereo microscope and images were acquired with the Leica DC500 digital camera.

3.9 Recovery of full length

The entire sequence of each isogroup of interest was recovered by RACE method (Rapid Amplification of cDNA Ends) using the same primers designed for qRT-PCR (Figure 3.2). A single strand full length cDNA library was generated from total RNA of a pool sample using SMART technology according to the protocol previously described (Chapter 3.4.1). To increase the annealing temperature and therefore the specificity of PCR, we designed a new set of primers characterized by higher melting temperatures. We replaced T7 primer with the longer Primer II A and designed a new Oligo$(dT)_{20}$-IIA for the first strand retro-transcription. Furthermore, a longer T3 primer, named T3Long, was designed. RACE reaction was performed using Advantage 2 PCR Polymerase kit (Clontech) in a 50 µL reaction volume according to the manufacturer's protocol. 5'-end was amplified with the T3Long forward primer and the specific reverse primer; instead, to recover 3'-end we used the specific forward primer and the Primer II A. The thermal cycling condition were as follows: 2 minutes of denaturation at 95°C, followed by 35 cycles of 15 seconds denaturation at 95°C, 20 seconds of annealing at 60°C, 2 minute of elongation at 72°C, and a final 5 minutes elongation at 72°C. Resulting amplicons were assessed by gel electrophoresis (1% agarose) and cloned using T4 DNA ligase (NEB) and pCR 2.1 vector (Invitrogen). Electro competent *E.coli* (DH10b) were transformed with pCR 2.1 recombinant vector and several colonies were screened using PCR and verified by Sanger sequencing (BMR genomics).

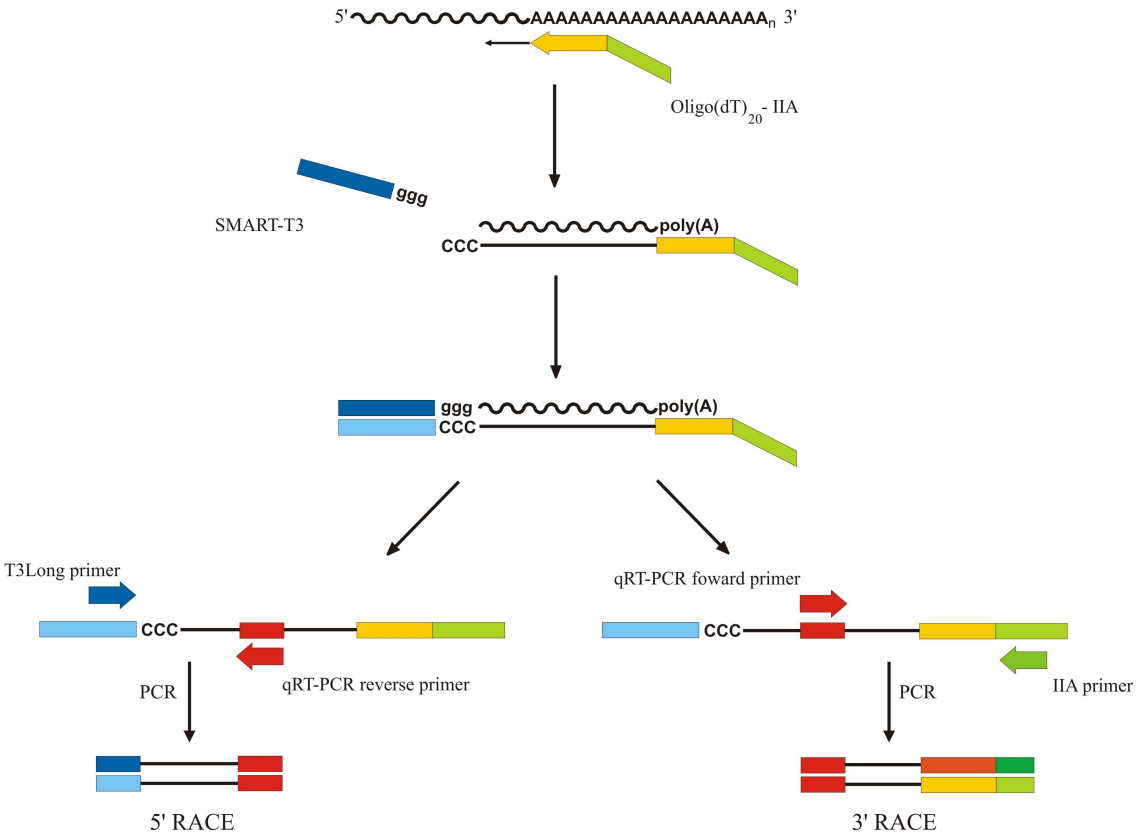

Figure 3.2 : Schematic representation of the 5' and 3'-RACE protocol. After SMART reaction we are able to selectively amplify the 5'-end as well as the 3'-end using the primers pair designed for qRT-PCR assay.

4 Results

4.1 Experimental design

The aim of this work is to define the transcriptional signatures of the 6 larval stages of *Mytilus galloprovincialis.*

First of all, sample of each larval stage had to be collected. In their natural habitat mussels shed eggs and sperm into open water where fertilization takes place. Mussel larval period is spent in the plankton, swimming and floating up in the water with a lot of other animals, plants, archaea, and bacteria. From a technical point of view it was impossible to collect species and stage specific samples from the plankton due to its heterogeneous composition. Therefore, we decided to perform mating, larval development, and sampling in laboratory under controlled condition, and using filtered water to avoid contamination from other organisms.

Microarray experiments are the most used approach to study genome-wide gene expression. Unfortunately little is known about *M. galloprovincialis;* its genome has not been sequenced and less than a half of expected transcripts have been identified. The sequences deposited on MytiBase result only from adult samples, so they do not include transcripts expressed only during larval development. So we are not able to design a microarray platform useful to define larval expression profiles. We decided to use the rising next-generation sequencing technologies to identify larval-specific transcripts. Between the available sequencing platforms, we used 454 GS FLX Titanium technology because, without a reference genome, the 300 nucleotides sequences produced make assembly and annotation process more simple.

We decided to set up and sequence seven stage specific not normalized cDNA libraries in order to identify specific larval sequences and obtain an estimation of their expression level by counting their representation frequency in each library. Therefore, we have developed a new protocol to construct a library of 3'-terminal cDNA fragments characterized by an optimized length for 454 sequencing (300-800 nt). By focusing the sequencing only on the 3' terminal region, each transcript gives only one sequence; so the number of produced sequences by each gene depends only on its expression level and it is independent of the transcript length. This approach allows us to perform a direct counts of transcripts abundance; without the need for estimates and

normalizations.

To validate the results of 3'-end sequencing approach we had to use an independent technique. Quantitative RT-PCR was performed to verify expression level defined by sequencing reads count in samples from a new mating experiment in order to increase the robustness of our results. Finally we performed *in situ* hybridization to define the morphological localization of several differentially expressed genes and also to start the functional characterization of some unknown genes.

4.2 Collection of mussel specific developmental stage

To set up a mussel mating experiment, not only specific equipments but also experience is required. So, we decided to collaborate with the *Istituto Delta,* a spin-off of University of Ferrara that is involved in the preservation of biodiversity along the coasts of the Adriatic Sea, in development of new aquaculture techniques, and in induction of spawning and rearing of bivalve larvae. In natural conditions, spawning may be triggered by several environmental factors including temperature, chemical and physical stimuli, water currents or a combination of these factors. There are several approach to induce a synchronous gametes release including biological methods, chemical stimulation and physical shock (Aji LP, 2011). Biological methods are based on the observation that sperms in the water frequently triggers spawning in other animals of the same species. Physical methods are able to mimic during few hours the long-range seasonal fluctuations in salinity or temperature that are important to the synchronization of the gametogenic cycle. Finally, chemical stimulation include injection of serotonin or sex steroids, but these approaches are difficult to be carried out on small bivalves such as *Mytilus galloprovincialis*. Moreover, all these approaches are not very effective to induce spawning of *Mytilus* genus. On the basis of a protocol set up by Dr. Edoardo Turolla (Istituto Delta, Goro), the best way to induce spawning of *Mytilus galloprovincialis* is based on the water treatment with a small quantity of hydrogen peroxide for few minutes. In fact, variation of water oxidative power has been reported to induce spawning in sexually mature bivalve by activation of synthesis of several prostaglandins naturally involved in spawning behavior of many different marine organisms (Morse DE, 1977). The natural period of spawning in *M. galloprovincialis* differs with geographic location: in Adriatic sea, it is usually confined between October and May with two peaks of emission at December and February. We decided to perform

the first mating experiment at February 2009. 20 males and 8 females have spawned within 1 hour after induction with H_2O_2 and they were selected to be used as parents in the crossing experiment. The difference between numbers of males and females is in agreement to the natural female/male ratio of 1/3. Rearing lasted 35 days and the timing of sampling is summarized in Table 4.1. Using these samples we constructed seven stage specific 3'-end cDNA libraries for sequencing. At February 2011 we decided to repeat mussel spawning induction, this new experiment was performed to validate sequencing results on a new biological sampling. In this second experiment, the same number of males (20) and females (8) were crossed to reduce the experimental variability. We collected 4 sample of each larval stage in order to extract total RNA and proteins and to fix larvae for morphological characterization and *in situ* hybridization. In addition to larval stages, we also collected eggs, sperms, embryos (1 hour after fertilization) and juvenile mussel (1 week after metamorphosis) to have a more complete overview of the developmental process in *Mytilus galloprovincialis*.

samples	timing
embryons	1 h
trocophore	1 d
D-larva	3 d
veliger	8 d
umbo-stage	14 d
pediveliger	19 d
metamorphic larva	25 d

Table 4.1 : Comparison between timing of the two samplings.

4.3 Total RNA extraction and quality control

We extracted total RNA from all samples using TRIzol reagent. Spectrophotometric analysis detected 260/230 ratios lower than 1.8 indicating a carbohydrate contamination. This is a common problem dealing with molluscs because they are rich in glycogen. So, we decided to perform a lithium chloride precipitation of that allows the precipitation of RNA in water without the addition of polar compounds, such as ethanol, that limits carbohydrate solubility. Therefore, glycogen remains in solution and does not precipitate with the RNA. We obtained about 50 µg of total RNA from each larval stage without carbohydrate or protein contaminations. Then, the integrity of extracted RNA was verified using capillary electrophoresis (Agilent Bioanalyzer 2100).

Usually RNA integrity is assessed by estimating the ratio between signals of 28S rRNA and 18S rRNA; a optimal 28S:18S ratio is 2.7:1. In several organisms, 28S is characterized by high instability resulting in 28S:18S ratios lower than the optimal value. The molecular mechanism for this type of instability is poorly understood, but it may result from the presence of breaking point of secondary and tertiary rRNA structure. This feature is particularly evident in the electropherograms of mussel RNA (Figure 4.1), where the 28S peak is almost absent. However, the good quality of 18S peak and the absence of fragments of low molecular weight suggest the absence of large-scale degradation of the RNA sample. Moreover, the absence of peaks of high molecular weight excludes genomic DNA contamination and confirms the quality of extracted total RNA.

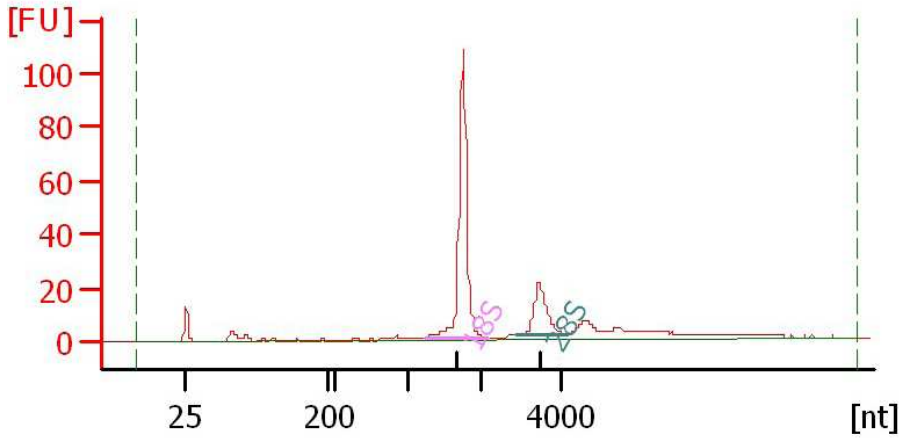

Figure 4.1 : Electropherogram of extracted total RNA. Total RNA extracted from D-Larva stage and analyzed with the 2100 Agilent Bioanalyzer using a RNA 6000 Nano LabChip. The high quality of total RNA is confirmed by the presence of intact 18S ribosomal peak (no RNA degradation) and no additional high weight peaks (no DNA contamination).

4.4 Setting up of the 3'-end cDNA libraries

We developed a protocol to construct 3'-end cDNA libraries compatible with 454 GS FLX Titanium sequencing technology. Key point of the protocol is the addition of a biotin molecule to the 3' end of each retro-transcribed first-strand cDNA to recover only its 3'-UTR after the fragmentation step.

Our protocol is characterized by 5 steps (Figure 4.2):

1) First-strand cDNA synthesis using SMART Technology;
2) cDNA amplification by PCR;
3) fragmentation by nebulization;
4) 3' fragments selection;
5) Poly(A) cleavage.

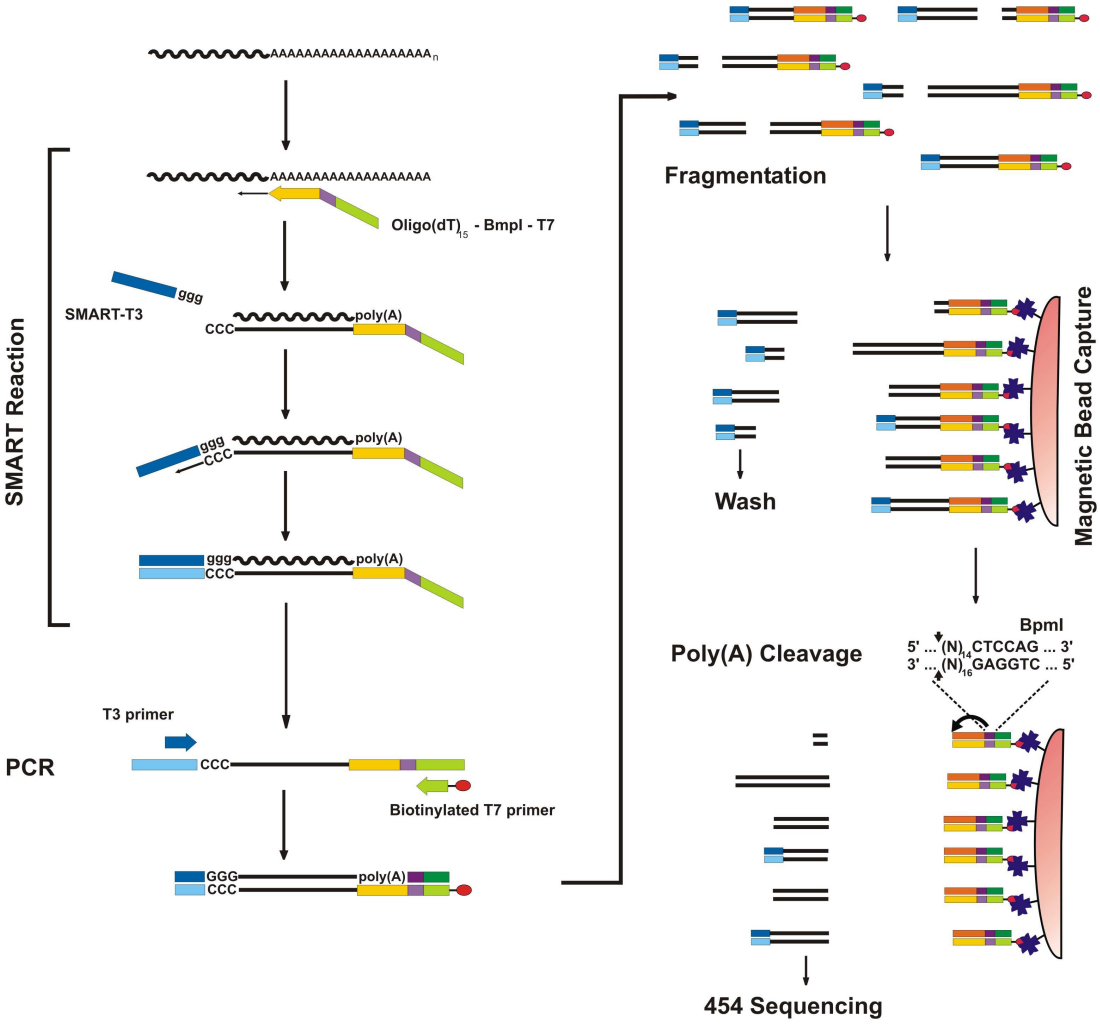

Figure 4.2 : Schematic representation of the protocol to construct the 3'-end cDNA library.

4.4.1 First strand cDNA synthesis

SMART (Switching Mechanism at 5' End of RNA Template) technology allows to insert known sequences at both ends of cDNA during first strand synthesis. A modified oligo(dT) primer was used for the first-strand synthesis from the poly(A). When M-MLV reverse transcriptase (RT) reaches the 5'-end of the mRNA, the terminal transferase activity adds few additional deoxycytidines (dCTP) to the growing cDNA. The SMART primer contains three riboguanosines that base-pairing to the deoxyribocytidine stretch, creating an extended template. M-MLV RT switches template and continues to the end (Chenchik *et al.*, 1998).

We used a peculiar oligo(dT)$_{15}$-BpmI-T7 primer that consists of: an oligo(dT)$_{15}$ with a degenerated base at the 3'-end allowing to start the retro-transcription from the first fifteen adenosines of the poly(A); a recognition site for the endonuclease BpmI to

release the 3'-fragments from magnetic beads after selection; and a T7 sequence useful to the second strand synthesis by PCR. The SMART primer consists of a T3 sequence, used as forward primer during amplification, and three riboguanosines at the 3'-end to bind the deoxyribocytidines added by the retro-transcriptase taking advantage from the greater power of bind that characterizes RNA/DNA base-pairing. Moreover, to increase the annealing between the neo-synthesized cDNA and the primer we used a SMART primer concentration five times higher than in a normal retro-transcription. To assess SMART reaction result, we performed a capillary electrophoresis analysis. The electropherogram of the first-strand cDNA, reported in Figure 4.3.C, shows a optimal cDNA length distribution up to 4000 nucleotides and it is characterized by the absence of low molecular weight fragments suggesting a good integrity of the first-strand cDNA.

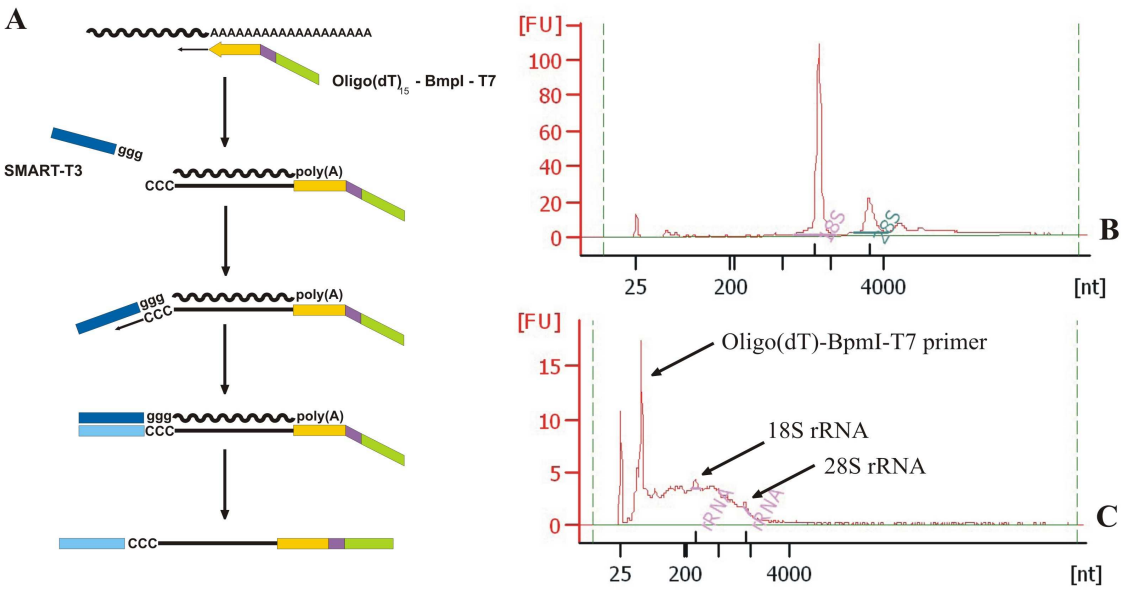

Figure 4.3 A : Schematic representation of first-strand cDNA synthesis with SMART technology. A modified oligo(dT) primer starts the first-strand synthesis. When M-MLV reverse transcriptase reaches the 5'-end of the mRNA, terminal transferase activity adds few additional deoxycytidines to the growing cDNA. The annealing of SMART primer produces an extended template. Reverse transcriptase switches templates and continues to the end of the transcript. T3 is indicated in blue, T7 in green, oligo(dT)$_{15}$ in yellow, and BpmI recognition site in purple. **B : Electropherogram of total RNA.** Starting total RNA extracted from D-larva stage and analyzed with the 2100 Agilent Bioanalyzer using a RNA 6000 Nano LabChip. **C : Electropherogram of first strand cDNA.** First-strand cDNA obtained after SMART reaction. The good distribution and the absence of accumulation of low molecular weight fragments demontrate the quality of cDNA obtained. Peaks of modified oligo(dT), 18S and 28S are also visible.

4.4.2 PCR amplification of the ssDNA

After SMART technology, we have performed a PCR in order to synthesize the second cDNA strand and to produce a quantity sufficient for next protocol steps. We have amplified al ssDNA sequences using the universal T3 and T7 primers linked to the 5' and 3' ends respectively. In this way, only full length first strand cDNAs were successfully amplified. We used this amplification step to lable the 3'-end of each ds cDNA with a biotin molecule (5' biotinilated T7 primer). The PCR was performed with a proofreading Taq polymerase to avoid nucleotide-misincorporation errors, especially harmful to applications where sequence accuracy is crucial, such as sequencing.

We have decided to block PCR during the exponential phase at 21 cycles to avoid distortions of the actual concentration of mRNA in the sample (Innis MA, 1990). We have demonstrated that PCR cycles between 19 and 22 are sufficient to achieve a good level of amplification without reaching the plateau phase in which the expression differences would be invalidated (Figure 4.4.A). To check the quality of sequeces produced with this protocol we have cloned and sequences some insert of D-larva cDNA library eith Sanger technology. As you can see in Figure 4.4.B, all sequenced inserts showed both T3 and T7 sequences that we have added during the first-strand reaction with SMART technology.

4.4.3 Fragmentation of double-strand cDNA

Next generation sequencing platforms produce small reads, so a fragmentation step is necessary to guarantee a complete coverage of transcripts of interest. Fragments length must not exceed twice the typical length of reads to avoid a lack of coverage and must be at least the median read length to take advantage of the full potential of the instrument. 454 GS FLX Titanium sequencing technology produces reads of 400 nucleotides with a median length of about 300 nucleotides, for this reason manufacturer recommends to sequence fragments ranged between 300 and 800 nucleotides to maximize yield of sequencing. Gel electrophoresis results (Figure 4.4.A) and capillary electrophoresis analysis (Figure 4.4.C) showed that cDNA libraries length distribution ranged from 400 to about 3500 nucleotides. We have tested some of the most used fragmentation methods to find the best approach. We have preliminary discarded some approaches: a) sonication damages the ring structure of DNA sugars decreasing efficiency of sequencing adapters ligation; b) hydrodynamic shearing needs for expensive equipments; and c) DNaseI approach needs for a DNA repair step after treatment.

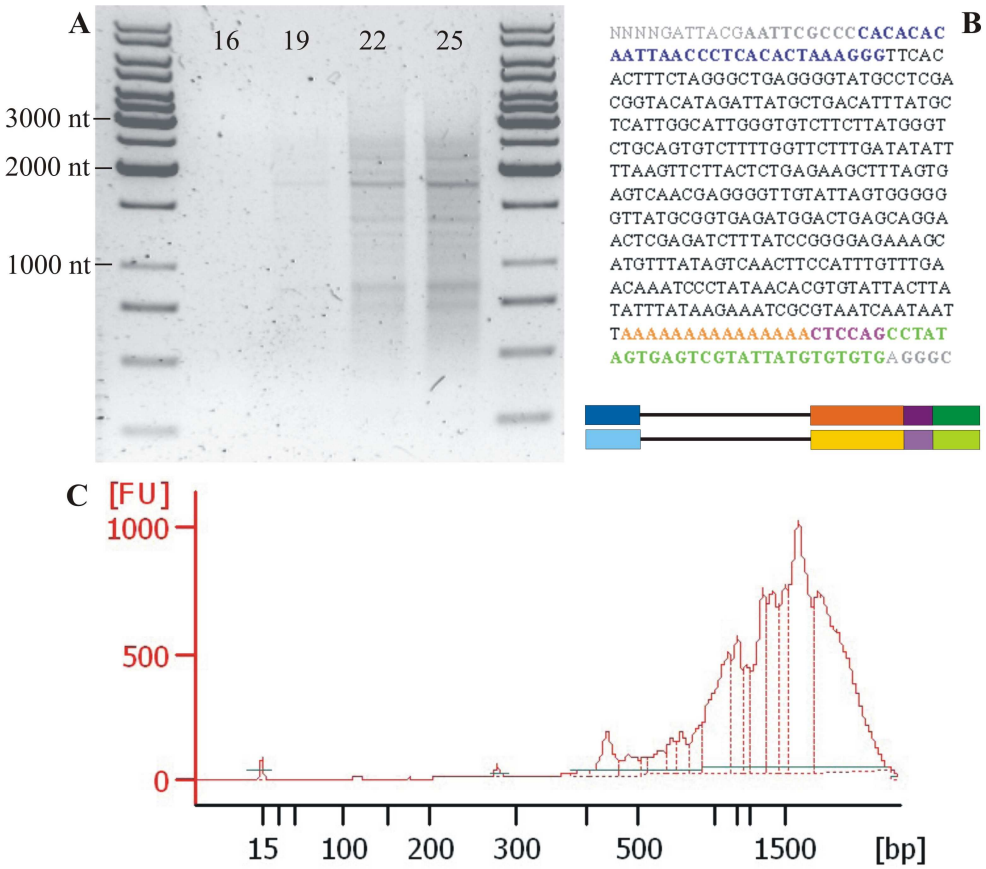

Figure 4.4 : Second-strand synthesis by PCR for D-larva library. A : Determination of the optimal PCR amplification cycle. The 1% agarose gel shows the products of D-larva stage first-strand cDNA amplification reaction stopped at 16, 19, 22, and 25 PCR cycles. **B : Sequence attesting the correct functioning of SMART reaction.** An example of sequence obtained from D-larva stage cDNA library showing the entire sequence of the modified oligo(dT) primer (indicated in orange, purple and green) as well as the sequence of SMART primer (indicated in blue). Grey bases are vector sequences. **C : Electropherogram of amplified cDNA.** Amplified and labelled cDNA obtained after SMART reaction and PCR. Library length distribution ranges from 400 to about 4000 nucleotides.

We have focus our attenction on: enzymatic digestion, temperature shock, and nebulization.

Temperature shock: this approach is based on RNA partial degradation with heat shock in a high salt concentration buffer. We tested different temperatures from 74 to 80°C for 2 minutes. As you can see in Figure 4.5.B we have obtained 200 nucleotides small fragments that are too low respect the manufacturer's instructions.

Enzymatic digestion: endonuclease digestion is the simplest method, but the resulting fragmentation is not random and does not result in a collection of overlapping fragments. For this reason, we have performed independent enzymatic digestions with six different endonuclease and we have pooled together resulting fragments. An enzyme with a 4 bases recognition site approximately cuts double-strand DNA every 256 base

pairs; instead, an enzyme with 5 nucleotides recognition site cuts every 1024 base pairs. In order to obtain about 500 bp fragments, we have decided to use 3 different enzymes with a 5 bases recognition site with a degenerated base (such as BstNI: $CC^{A}/_{T}GG$) cutting dsDNA every 512 nucleotides, and 3 enzymes with a 4 bases site composed only of C and G nucleotides (such as HhaI: GCGC) that are less frequent dNTPs. No good results were obtained from this approach (Figure 4.5.C) because the majority of longer sequences were not digested by the enzymes. So, we have decided to discard this method.

Nebulization: the double-strand DNA fragmentation occurs by forcing a DNA solution through a small hole. The size of fragments is determined by the speed at which the DNA solution passes through the hole. Pressure of nitrogen gas through the nebulizer, viscosity of nebulization buffer, and temperature are important parameter to determine double-stand DNA speed. We have decided to modify only the gas pressure and the exposure time. We tested 30, 35, 40, and 45 psi and we identify 30 psi as optimal pressure. As you can see in Figure 4.5.E, an optimal distribution of fragments was obtained. The main disadvantages of this method are represented by the large amount of starting material and the extraction of nebulized cDNA from the nebulization buffer that is rich of glycerol. These problems were easily solved because the amplification step produced high amount of starting material and we have perform a simple precipitation to discard nebulization buffer and concentrate cDNA.

4.4.4 Selection of 3’-end cDNA fragments

After nebulization only the biotinilated 3’-end fragments were recovered using magnetic beads conjugated with streptavidin, while other cDNA fragments were discarded with abundant washes. We have used an excess of beads in order to ensure the binding of all 3’-end fragments. Bead capture was performed in a high volume of binding buffer and several washes were performed to discard all not 3’-end fragments and reagents of the previous protocol steps.

4.4.5 Recovery of 3’-end cDNA fragments from magnetic beads

Since, the biotin molecule could make the sequencing adapters ligation difficult, fragments were released from magnetic beads by endonucleolytic digestion of modified oligo(dT) used to prime SMART reaction (Figure 4.6.A). Enzymatic digestion allowed to remove biotin molecule, and to eliminate poly(A) stretch (15 nt) that negatively affects the sequencing.

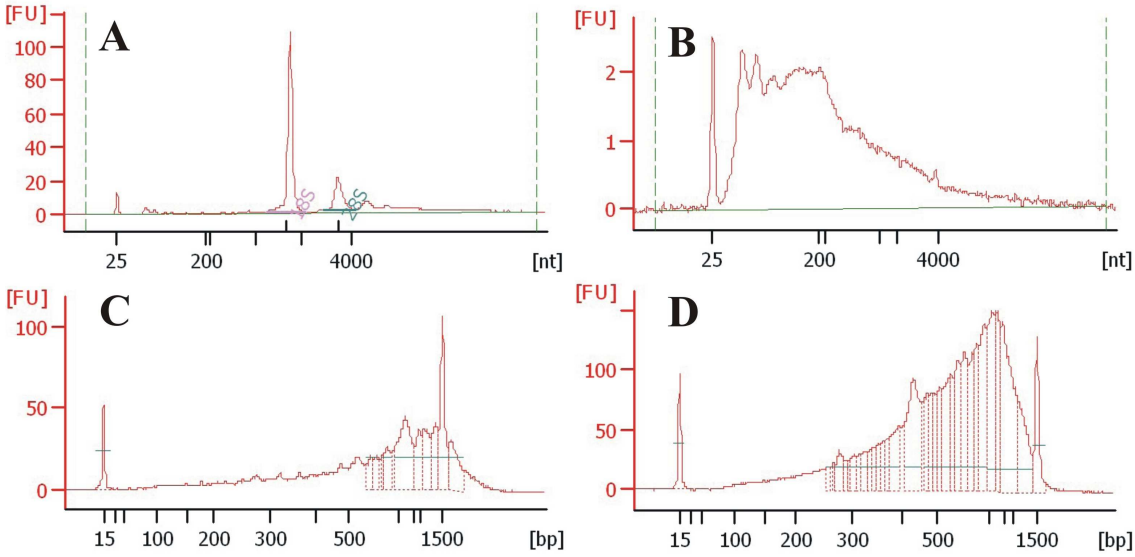

Figure 4.5 : Qualitative analysis of fragmentation protocols. All the Agilent Bioanalyzer electropherograms are referred to various phases of construction of D-Larva stage cDNA library. **A : Starting total RNA extracted from D-Larva stage. B : Fragmentation by temperature shock method.** Electropherogram of total RNA incubated at 74° for 2 minutes. A large quantity of small RNA fragments (< 200 nucleotides) was observed. **C : Fragmentation by enzymatic digestion.** Electropherogram of dsDNA digested with six different restriction endonucleases**.** Only a small fraction of long dsDNA were digested. **D : Fragmentation by nebulization.** Electropherogram of nebulized dsDNA at 30 psi of nitrogen pressure. A complete fragmentation of all dsDNA sequences larger than 1500 base pairs was obtained. We have not observed an increase in the abundance of small fragments shorter than 300 nucleotides. RNA and cDNA samples were analyzed with the 2100 Agilent Bioanalyzer using a RNA 6000 Nano LabChip and a DNA 1000 LabChip respectively.

The major limitation of 454 sequencing relates to homopolymers because, nucleotides used in pyrosequencing reaction, have not terminating groups preventing multiple nucleotides incorporation at each cycle. The incorporation of two consecutive identical bases results in a double intensity fluorescent signal. However, the homopolymers signal is linear only up to eight consecutive nucleotides after which become difficult to infer the original stretch length, and a significant number of reads are filtered out during quality analysis process. For this reason, in trascrptome sequencing a lot of reads are discarded because most of cDNA library protocols use oligo(dT) to start retro-transcription so the large majority of sequences presents an homopolymer at the 3'-end. To solve this problem we have adopted a method published by Frias-Lopez *et al.* (Frias-Lopez, 2008) based on BpmI that is a SII class restriction endonuclease. We have inserted a BpmI recognition site in the oligo(dT)-T7 primer used in the SMART

reaction to product the first strand of cDNA. BpmI cleaves 16 base pairs downstream its recognition site eliminating the entire poly(A).

As you can see in Figure 4.6.B, the BpmI digestion does not alter length distribution of fragments. This result demonstrates that BpmI treatment did not lead to cDNA degradation.

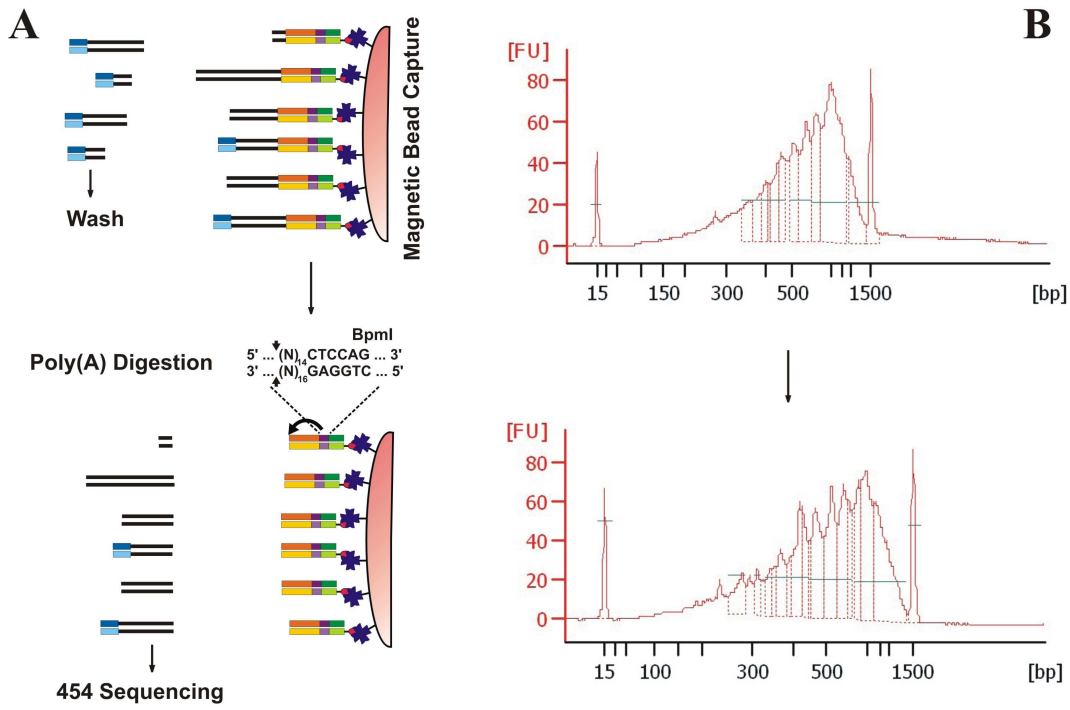

Figure 4.6 : Construction of the 3'-end cDNA library. A : Selection of 3'-end cDNA fragments. Double-strand cDNA molecules were digested with BpmI in order to break the link with magnetic beads and to eliminate poly(A) recovering the 3'-end cDNA fragments. **B : cDNA library length distribution before and after BpmI digestion.** Up and down electropherograms represent the fragments length distribution of the D-larva cDNA library before and after BpmI digestion respectively. We did not observe difference in length distribution after BpmI digestion demonstrating the success of the protocol. Electropherograms were obtained with a DNA 1000 LabChip.

4.4.6 Quality and quantity analysis of cDNA libraries

To assess the quality of resulting libraries we used three different technologies. First of all, we used Nanodrop to quantify the library and to verify the presence of carbohydrates and protein contamination. Then we performed an Agilent Bioanalyzer assay with DNA 1000 chip to verify the correct length distribution of 3'-end cDNA fragments. Finally we repeated cDNA quantification using Qubit fluorometer that allows to quantify specifically the type of nucleic acid obtaining a more reliable quantification compared to NanoDrop because this quantification is not affected by the presence of nucleotides and single-stranded nucleic acids. Comparing the two

quantification results it is possible to determine the purity degree of each sample.
We have obtained from 1 to 1.6 µg of double strand cDNA from each library. The difference between NanoDrop and Qubit quantification never exceeds the 15%; this observation and the good 260/230 and 260/280 ratios demontrate the high purity of libraries.

4.5 Next generation sequencing of larval specific cDNA libraries

Sequencing of the seven stage specific 3'-end cDNA libraries was performed with 454 GS FLX Titanium technology, according with the manufacturer protocol at BMR Genomics S.r.l.. We decided to use an eighth of 454 plate for each library. 454 sequencing generated about 1.2 millions of reads: 11.5% were discarded by quality filters and another 25.7% because too short reads. Remaining 751,872 high quality reads (62.8%) showed a median length of 315 nucleotides (Figure 4.7). A list of sequencing results for each cDNA library is presented in Table 4.2.
On the basis of the instrument yield suggested by the manufacturer - about 70,000 reads for line and 300 nt of median length - we obtained a good sequencing run. In particular, the number of reads generated is about 50% higher than the expected and six of the seven cDNA libraries are characterized by a median length higher than expected 300 nt.

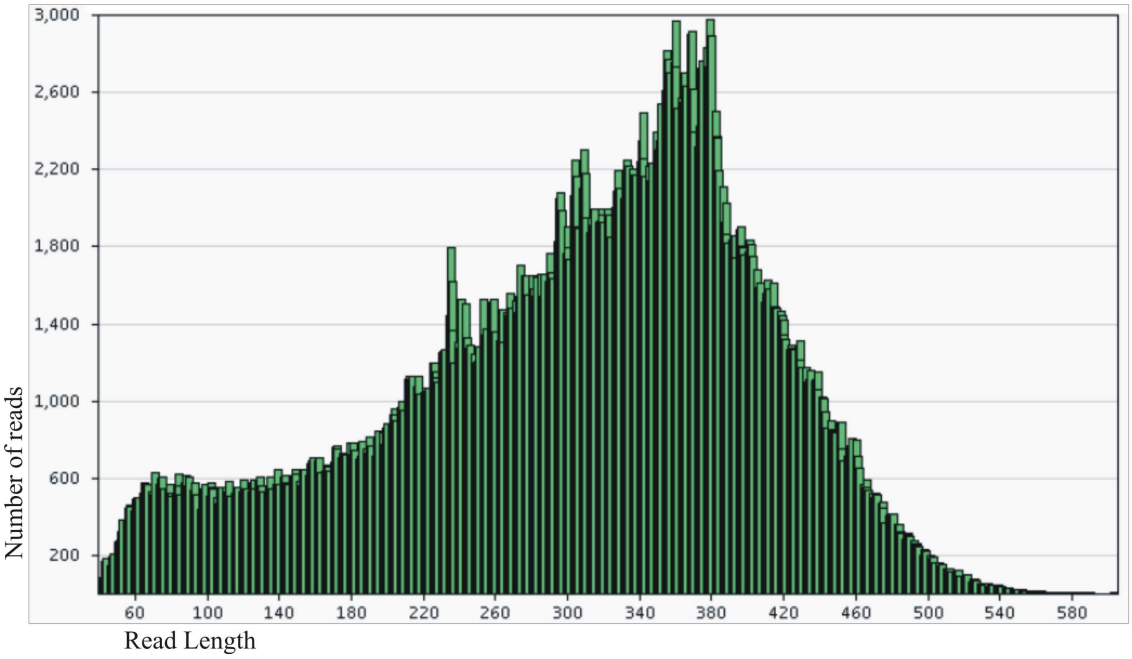

Figure 4.7 : Read length distribution of 454 sequencing run. Length distribution of all the 751,872 high quality reads produced by the sequencing of the seven libraries. Length profile agrees with the expected optimal distribution.

Larval Stage	Number of reads	Median reads length
embyos	103,009	326
trocophore	95,664	281
D-larva	84,554	303
veliger	130,098	334
umbo-stage	101,863	334
pediveliger	127,855	305
metamorphic	108,829	310
TOTAL	751,872	315

Table 4.2 : 454 sequencing results of each stage-specific cDNA library.

4.6 Analysis of 454 sequencing data to verify the library protocol

Before the assembling of 454 reads we decided to understand if our library protocol worked properly. To this aim 454 reads were aligned to known *M. galloprovincialis* ESTs deposited on Mytibase with stringent parameters: 95% of identity and at least 90% of coverage. Since MytiBase sequences were obtained only from adult samples, we were able to compare only reads of transcripts expressed from larval to adult mussel life. Only 2723 of the 7112 (38%) MytiBase sequences were aligned to 454 reads, attesting an high gene discovery rate associated to stage-specific cDNA libraries.

In Figure 4.8 we reported some example of transcripts characterized by different lengths. For 400-700 nt transcripts we were able to obtain the entire sequence because sequencing reaction started from 3' or 5'-end of cDNA fragment producing reads of 300-400 nt that are able to cover the entire sequence. Whereas, only the 3'-end region was sequenced for transcripts longer than about 800 nt, demonstrating the good functioning of our protocol.

To evaluate the fragmentation frequency, we calculated the ratio between the number of 5' and 3'-end reads that identify the same transcript. We can observe about 1:1 ratio for the smaller transcript (300 nt) indicating a low fragmentation frequency, 2:5 ratio for the 700 nt transcript, and 1:244 ratio for 1,200 nt transcript attesting an high fragmentation frequency. Therefore, fragmentation frequency is directly proportional to cDNA length and 454 sequencing of our libraries has produced 3'-end reads for all transcripts and 5'-end reads only for transcripts up to 800 nt long.

We have also counted the number of reads that present the modified oligo(dT) primer sequence in order to evaluate the efficiency of BpmI digestion step. Less than the 0.0001% of reads showed this sequence, demonstrating the success of enzymatic reaction

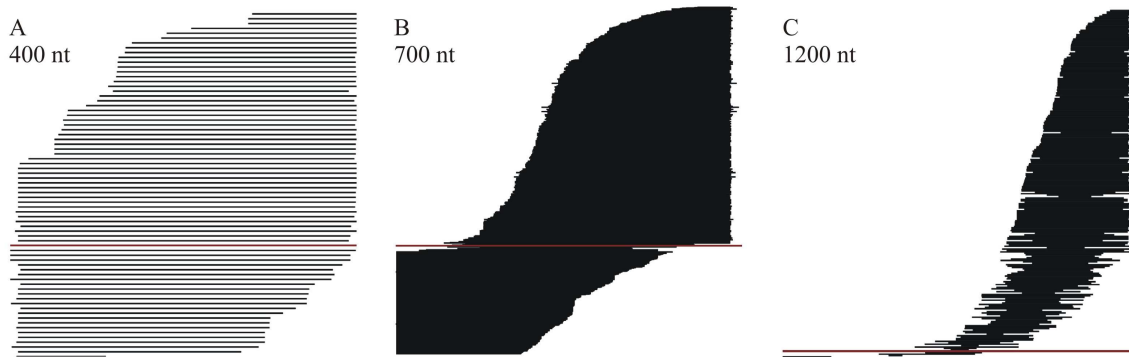

Figure 4.8 : Alignments of 454 reads and Mytibase consensus sequences. Black lines represent 454 reads generated from transcripts with different lengths: 400 nt (**A**), 700 nt (**B**), and 1,200 nt (**C**). Red lines consensus sequences deposited on MytiBase.

Analyzing alignments between 454 reads and MytiBase consensus sequences we noticed some unexpected cDNA processing modifications.

Inner poly(A): some transcripts showed inner adenosine stretches that act as primers for retro-transcription reaction (Figure 4.9.A and B). If two oligo(dT) bind a transcript at the same time, the unespected retro-transcription, started from the inner adenosine stretch, blocks the first-strand synthesis from the poly(A), preventing it to accomplish SMART reaction. Since the annealing between the modified oligo(dT) and adenosine stretch is proportional to the length of the homopolymer, if the adenosine stretch is small (< 10 nt) we are able to recover both sequences (from 3'-end and from inner stretch) (Figure 4.9.A); while, if the adenosine stretch is longer (Figure 4.9.B) we lose information about the 3'end region. This problem did not affect the reads count because the key point of our experimental design - one transcript gives only one read - is still valid.

Constitutive BpmI recognition site: the BpmI recognition site is a six nucleotides long sequence that statistically cut DNA every 4096 nucleotides. Some transcripts could have one or more BpmI recognition sites in their sequence generating different fragments after BpmI digestion. Therefore, the same transcript could be identified by several not overlapping fragments that generate independent contigs. Fortunately, BpmI digestion seems to be characterized by a 99% cutting efficiency. If a transcript is highly expressed probably will generate not cut sequences that can be used to join the contigs (Figure 4.9.C); instead, the probability to obtain overlapping reads for low abundance transcripts is low (Figure 4.9.D). Moreover, if one of the two fragments, generated by BpmI digestion, is less than 300 nt the probability to lose it is very high during purification steps before emulsion-PCR in the sequencing process (Figure 4.9.E). The

presence of a BpmI recognition site affect the quantification of gene expression levels because the same transcript is identified by two fragment and its representation frequency doubles. Comparing expression levels of a transcript across stages, BpmI activity does not affect the analysis because the transcript is processed in the same way in all libraries. But, if we want to compare expression level of two different transcripts we have to take into account this problem. We have developed an algorithm to count the number of reads that could be digested by BpmI in order to estimate the magnitude of the problem. We decided to count the number of reads that showed a BpmI recognition site localized near the end. About the 16% of reads seems to be fragments generated by BpmI digestion. We repeated this analysis on the assembled genes and singletons (Chapter 4.7) to check if assembler software was able to use overlapping sequences to join the segregated contigs. Up to the 8% of genes and singletons could be still affected by this problem. Therefore, assembly step was able to rejoin only the 50% of cut sequences. To completely resolve this problem we would need larval reads obtained from a library not digested by BpmI. Without these sequences we have decided that the best way to face this problem was a manually correction. After annotation, for each isotig of interest we controled if there were one or more isotigs with the same description, and verified if they were fragments of the same transcripts. Moreover, before comparing expression level of two different transcripts, we examined the two sequences searching for BpmI recognition sites, and, if a transcript was cut, we verified its counts analyzing the alignment of the reads to estimate the real representation.

4.7 454 reads assembling and construction of a mussel development transcript catalogue

4.7.1 Assembling process

All 454 reads generated from the seven stage-specific libraries and 18,788 ESTs deposited on MytiBase were assembled together. We have tested and compared three different available software: Newbler 2.5.1 (Roche), CLC (www.clcbio.com/) and MIRA 3.2.1 (chevreux.org/projects_mira.html); results are presented in Table 4.3

After the comparative analysis we decided to use the assembler that demonstrated to be more reliable. CLC was discarded because only the 33% of starting 454 reads were assembled. MIRA classified a lot of sequences as outliers (too short or repeats), 3 times more than CLC and almost 20 times more than Newbler. But, this is not a real problem because all outliers were used by MIRA for the assembling process; the only difference

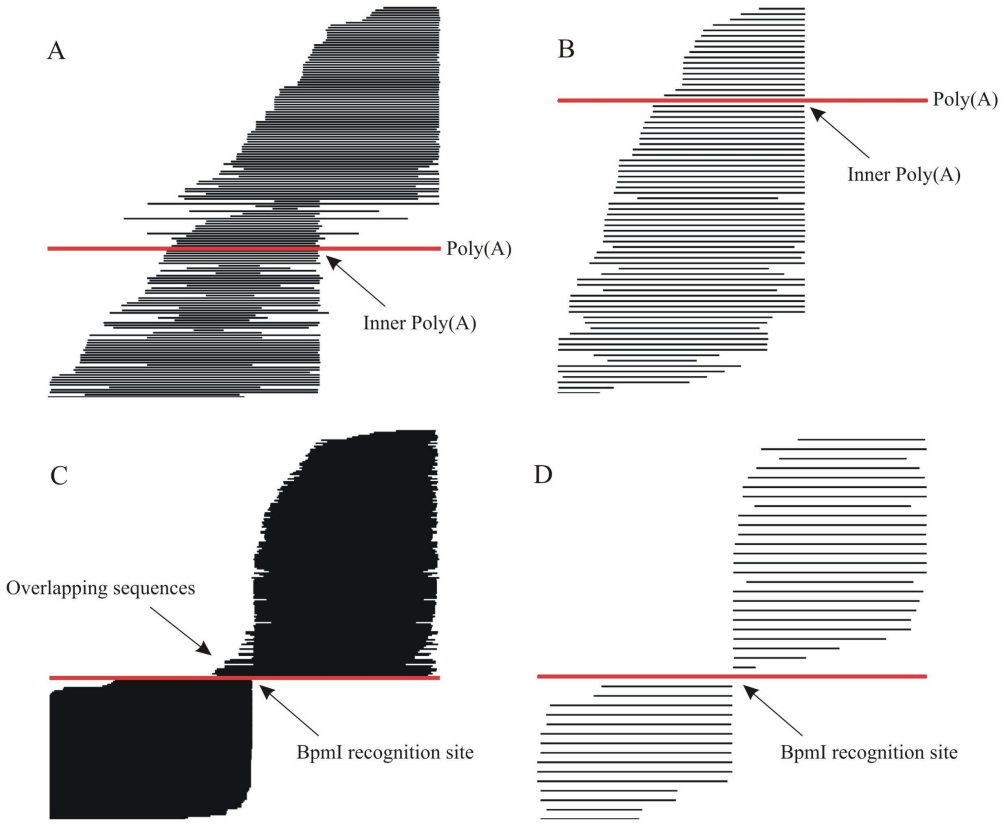

Figure 4.9 : Criticisms of the 3'-end cDNA library protocol. 454 aligned reads and MytiBase consensus sequences are represented by black and red lines respectively. All sequences are 5'→3' oriented. **A:** A canonical poly(A) and an adenosine stretch of 9 nt were associated to this transcript. **B :** Transcript with a inner adenosine stretch of 16 nucleotides. All generated reads started from the inner adenosine stretch. **C and D :** 2 examples of reads generated from sequences digested by BpmI restriction endonuclease. The first alignment shows an high expressed transcript characterized by some overlapping sequences (C); the second represents a low expressed transcript without overlapping reads (D).

between an outlier and a common read is that outlier can not be a starting sequence for alignment. Singletons (unique reads) are classified as "debris" and are discarded at the end of the process together with unused outliers. We expected to obtain singletons only from low abundance transcripts; so, 140 thousands singletons seems to be too much. Moreover, the number of isotigs/transcripts (53,000) is excessive for a mollusk with a genome of about 1.7 pg. For these reasons, MIRA 3.2.1 was discarded. Instead, Newbler 2.5.1 results to be a good choice: it was able to assembly an high percentage of 454 reads (83%) as well as MytiBase EST (72%), while the number of outliers is small (2%). Newbler results are characterized by long contig and relatively small number of singletons (10%). According to us, the number of singletons was yet excessive we decided to remap singletons on the isotigs just assembled with less stringent parameters in order to allow the pairing of a greater number of reads and to reduce the number of singletons. After remapping, we observed a 27% reduction of singletons. We believe

that such a large number of singletons could be due not only to the low expressed transcripts, but also to the random and unexpected sequencing of unspecific fragments from the 5'-portions of transcripts. Only singletons longer than 200 nucletides (37,471) were used for the following analysis.

	Parameters	**Newbler 2.5.1**	CLC	MIRA 3.2.1
454 reads	starting reads	**750,712**	751,629	748,562
	assembled reads	**655,835**	250,029	607,804
	singletons	**78,359**	405,806	NA
	outliers	**16,518**	95,794	307,863
MytiBase ESTs (18,788)	assembled ESTs	**13,568**	12,205	15,042
	singletons	**4,009**	6,583	NA
	outliers	**1,211**	0	3,746
Results	Isotigs	**14,364**	28,724	53,000
	Isogroups	**10,200**	28,724	NA
	Mean contig length	**559**	417	456
	Time taken	**4 h**	< 1 h	2 days

Table 4.3 : Comparison between three assembling software. Table reports the starting number of reads obtained after trimming and quality control performed by each assembler. The number of 454 reads and MytiBase ESTs assembled is reported as well as the number of sequences not aligned with any other (singleton) and the number of sequences classified as outliers because repeats or too short (> 20 nt). Finally, table reports the number of isotigs (putative transcripts) and isogroups (putative genes) generated, the mean contig length, and the time taken to accomplish the assembly.

4.7.2 Has 454 sequencing increased the transcriptome knowledge of *M. galloprovincialis*?

We have compared the sequences deposited on MytiBase with the new assembling data. 7112 unique transcribed sequences (2,446 consensus sequences and 4,666 singletons) were deposited on MytiBase; while, 51,785 unique transcribed sequences (14,364 isotigs and 37,471 singletons) resulted from Newbler 2.5.1 assembling. It is important to remember that 18,788 MytiBase ESTs were used as seed sequences in the 454 assembling process. In particular, 13,568 ESTs take part in 2,730 resulting isogroups; therefore, the assembling of Mytibase ESTs and 454 reads provides 284 consensus sequences more than the original MytiBase clustering. MytiBase (Figure 4.10.A) and new assembly (Figure 4.10.B) sequences length distributions were reported in Figure

4.10. 454 singletons was notreported in Figure 4.10.B because they are less too short altering length distribution, but also less reliable compared to the new consensus sequences and MytiBase Sanger singletons. Next generation sequencing approach has dramatically increased both the number and quality of the known transcripts. In fact, the number of sequences ranged from 200 to 800 nt were significantly increased, sequences between 800-1,000 nt were doubled and those longer than 1,000 nt were tripled.

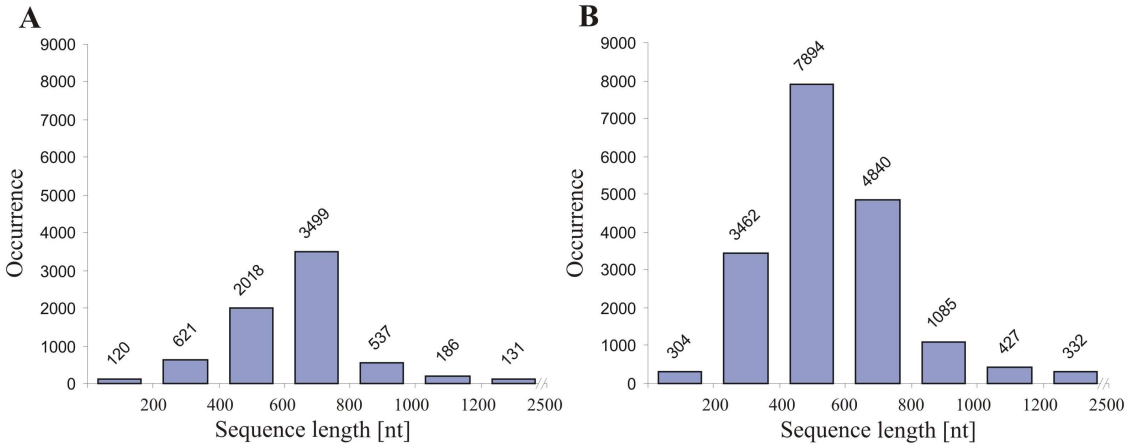

Figure 4.10 : Length distribution of assembled sequences. A : MytiBase sequences length distribution. Length distribution of the 2,446 consensus sequences and the 4,666 singletons deposited on MytiBase. **B : Length distribution of the new assembly dataset.** Length distribution of the 14,364 isotigs and the 4,009 singletons obtained after assembling between 454 reads and MytiBase ESTs.

4.7.3 Annotation process

Considering all orders in the class of Bivalvia, not a single genome has been fully sequenced yet. For this reason, annotation step usually results in a low percentage of annotated genes and most of the descriptions is derived from alignments with sequences of philogenetically distant organisms. For example, Huan *et al.* (Huan P, 2012) and Feng *et al.* (Feng B, 2010) annotated 12.2% and 20.7% of produced sequences for *Meretrix meretrix* and *Sinonovacula constricta* respectively. For these reasons we have decided to develop a customise annotation algorithm. Each non redundant sequence was searched in the nucleotides database (NCBI) and UniProtKB database using Blast-N and Blast-X with an e-value cut-off of e^{-50} and e^{-6} respectively. These values were empirically chosen considering the low amount of annotated sequences available for *M. galloprovincialis* and similar bivalve species, and the need for stringency in providing a reliable catalogue of Mediterranean mussel genes. To avoid wrong annotations, our algorithm has selected, among the best five descriptions, the annotation related to the organism characterized by the closer taxonomic distance to *M. galloprovincialis*. Using this approach we were able to annotate the 44.6% of isotigs (6,412), and focusing on the

differentially expressed isogrups (Chapter 4.8.2) this percentage rises to 48.8%. The higher percentage of annotation of differentially expressed genes is probably due to the availability of more information about genes involved in developmental process. It is interesting to note that differentially expressed transcripts of embryos (42.0%) and early larval stages (46.6%) are less annotated because less studied by scientific community. All annotations of altered transcripts were further manually examined, to verify if the best describing text was assigned to the corresponding cluster. About 13.7% of annotation were manually modified but in most cases (76.2%), correction was due to a bad or missing annotation of the reference sequence in public database. Our algorithm for automatic annotation was also able to correctly annotate specific proteins that share high conserved domains with common proteins. For example the program has correctly annotated shematrins family that shares functional domains with fibrinogen but is involved in the shell synthesis of several marine mollusks. Finally, our modified automatic annotation process gave high percentages of correct annotations, considering that *M. galloprovincialis* is a non-model organism.

These good results demonstrated the utility of 600-800 nt 3'-end sequences to guarantee an annotation efficiency comparable to that obtained by full length approaches.

4.8 Defining larval development expression signature

4.8.1 454 reads counting and cluster analysis

We evaluated the gene expression profile of each larval stage by counting the number of reads that align to each isogroup in the six stage-specific libraries. Since cDNA libraries have different reads abundances, results had to be normalized. We tested different normalization methods and verified results performing a hierarchical cluster analyses. First of all we performed a total reads normalization and a normalization using only the number of reads that were assembled for each library. Then, we tested a quantile normalization, usually used in microarray experiments, and a third quartile normalization. Surprisingly, regardless of the method of normalization chosen, and even using raw data, cluster analysis was able to arrange the stages in the correct temporal order from embryos to metamorphic larva. This observation was the first evidence of the reliability of our experimental approach.

Results are also independent of the clustering methods chosen; we tested Euclidean distance and Pearson correlation obtaining almost the same results. We chose

Euclidean distance because it groups elements not only on the basis of their common trend but taking also into account their differences of intensity. For example, Euclidean distance is able to better show the difference between the embryos library and the others. In fact, analyzing the counts obtained from each library, we noticed that embryos stage was characterized by several transcripts with expression levels hundreds of times higher than in the other stages. For this reason, we decided to individually analyze the embryos library.

In a paper published in 2010, Bullard *et al.* (Bullard X, 2010), compare several normalization methods for the analysis of quantitative RNA sequencing experiments and demonstrate that the third quartile normalization is the more robust statistical approach. According to these results we performed a third quartile normalization of the counts of the seven stage specific libraries. Cluster analysis (Figure 4.11) divided larval stages in two main groups: early larval stages including trocophore and D-larva, and late larval stages including umbo-stage, pediveliger, and metamorphic larva. Veliger belongs to late larval stages but seems to be a transition stage between early and late stages.

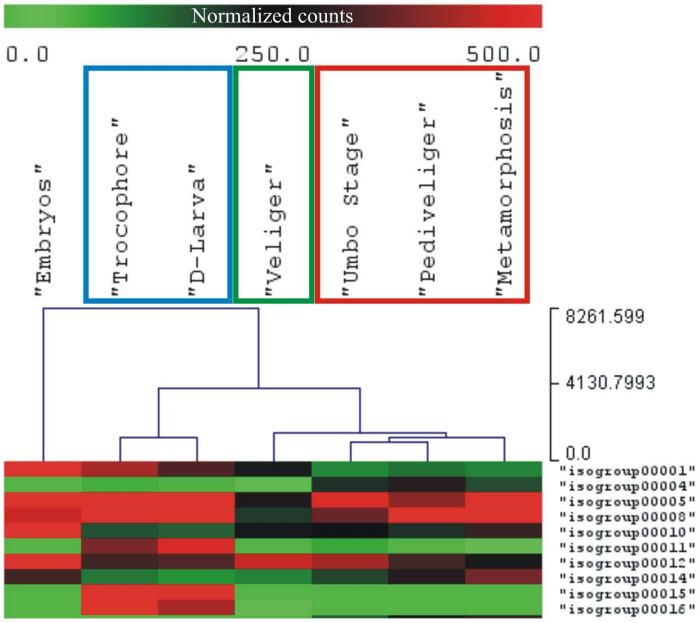

Figure 4.11 : Cluster analysis of normalized counts. Samples tree generated by hierarchical cluster analysis of the third quartile normalized counts using Euclidean distance. Cluster analysis is able to arrange the stages in the correct temporal order from embryos to metamorphic larva, and divides them in two groups: early larval stages (trocophore and D-larva; represented in the blue box) and late larval stages (umbo-stage, pediveliger, and metamorphic larva; represented in the red box). Veliger (represented in the green box) seems to be the transition stage while embryos stage differs from all the other stages.

4.8.2 Determination of stage-specific expressed genes in *M. galloprovincialis* development

The next step of the analysis was to define the significantly differentially expressed genes (Deg) across the larval stages. We could not use a statistical test based on a normal distribution because counts are discrete variable. We had to use a binomial test such as the Poisson distribution. However, Poisson model assumes that mean and variance are equal and this is not true for counts of 454 reads, and we did not perform experimental replicas of sequencing. So we used a modified Poisson distribution called binomial negative test, that estimates the variance from the total counts of each library.
A negative binomial test was performed on third quartile normalized counts to identify significantly differentially expressed genes between at least two stages. A binomial test makes a comparison between two categories at the same time, for this reason we performed all possible combinations comparing single stages or groups of similar stages or single stages *versus* groups. The 21.8% of isogroups (2,225) is differentially expressed in at least one larval stage. We decided to focus our attention on differentially expressed genes between larval samples clearly separated by hierarchical cluster analysis (see Table 4.4). The comparisons with higher number of Deg were those between early and late larval stages and between embryos and the other stages. Cluster analysis showed that veliger is more similar to late larval stages and this result is confirmed by the very low number of Deg (only 3) between veliger and late larval stages. However, veliger could be considered a transition stage because the number of Deg between veliger and early larval stages (257) is minor respect to the comparison between early and late larval stages (972). Also the number of Deg in early stage *vs* metamorphic larva comparison (503) was lower than in the early/late stages comparison. These observations suggested that there are two main groups of larval regulated genes, one peculiar of early stages and the other involved in umbo and pediveliger stages. Comparing trocophore and D-larva stages we did not find differentially expressed genes, so we decided to consider them as a single class during discussion of data. For the same reason also umbo-stage and pediveliger stage are grouped. Finally we noticed that the number of Deg in the comparisons between embryos and the other stages surprisingly decreased from early to late larval stages. These results confirm the peculiarity of embryos that showed a unique gene expression signature.

Comparisons between larval stages *vs*		**number of differentially expressed transcripts (Deg)**
embryos	early stages	1,101
embryos	veliger	989
embryos	late stages	758
early stages	veliger	257
early stages	late stages	972
early stages	metamorphic larva	503
veliger	late stages	3
umbo-stage + pediveliger	metamorphic larva	38

Table 4.4 : Number of differentially expressed genes from each comparison. For each comparison we have reported the number of Deg obtained with binomial negative test on the third quartile normalized counts.

4.9 Validation of differentially expressed genes by quantitative real-time PCR (qRT-PCR)

Quantitative RT-PCR analysis was performed to quantify and validate the expression level of some genes presenting different reads counting in specific mussel larval stages. We decided to exclude from this analysis the embryo stage because it showed a peculiar expression signature. First of all we had to find a suitable reference gene with constant expression levels in all the larval stages. The housekeeping genes most used in adult mussels are the 18S rRNA and actin. Several studies have tried to identify housekeeping genes in other organisms of the class bivalvia. Siah *et al*. (Siah A; 2008) demonstrating that the best housekeepings for *Mya arenaria* are 18S rRNA and Elongation Factor 1 α (EF1), while Morga *et al*. (Morga B; 2010) reported EF1 as the best choice for *Ostrea edulis*. Both these studies as well as our previous experiments on *M. galloprovincialis* rejected actin as good reference for adult samples. We decided to preliminary test the expression of 18S rRNA, EF1 and Actin at the same time comparing raw cycle thresholds of each stage. Ribosomal 18S resulted differentially expressed across larval stages; it was characterized by a stable expression profile during late stages but it was resulted over-expressed in early stages; also EF1 expression profile was not constant but showed an increase during umbo-stage. Instead, actin showed stable expression levels across larval stages resulting a good reference. Since these observation were in agreement to expression profiles obtained from reads counting of EF1 and actin (Figure 4.12; 18S rRNA counting profile is not represented because we have obtained few 454

reads), we have tried to identify the best housekeeping gene analyzing count expression profiles for transcripts with a constant expression level across larval stages. Between genes characterized by lower relative standard deviation we have found actin and some of the most well-known housekeeping genes such as calmodulin and clathrin. According to these results we concluded that reads counting could be helpful to identify putative housekeeping genesin non-model organisms. Validations were carried out on mussel samples collected during the second mussel mating experiment, performed at the beginning of 2011, to increase the robustness of our results. We selected 9 annotated differentially expressed genes and designed qRT-PCR primers specific to isotig sequences. In Figure 4.13 each graph reports the reads counting profile compared to qRT-PCR results with the associated Pearson correlation coefficient. Quantitative RT-PCR confirmed most significant expression differences among larval stages, but also demonstrated that our approach was able to accurately report even less marked differences. Finally, the expression values obtained with qRT-PCR for the tested transcripts were in agreement to reads counting in the seven stage-specific libraries as demonstrated by the high Pearson correlation values.

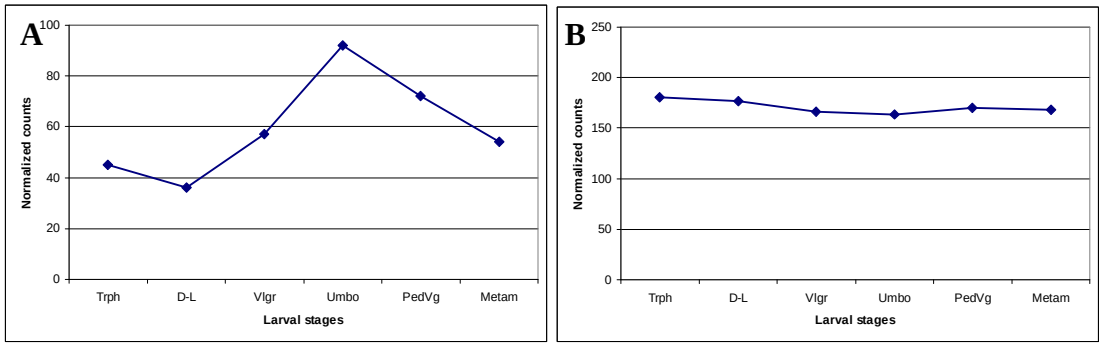

Figure 4.12 : Expression profiling of two candidate housekeeping genes. Expression profiles obtained from reads counting are useful to identify putative housekeeping genes. **A :** Elongation factor 1 is not a good reference because it shows a peak of expression during umbo-stage; instead, **B :** actin could be considered a good reference because it shows similar expression levels across larval stages.

4.10 Functional categorization of *M. galloprovincialis* stage-specific transcripts

Differentially expressed transcripts have been grouped into functional categories according to Gene Ontology (GO) in order to identify the biological process mainly involved in larval development. Our modified automatic annotation process selected the description more suitable for *M. galloprovincialis*, but, unfortunately descriptions referred to non model organisms are less GO annotated.

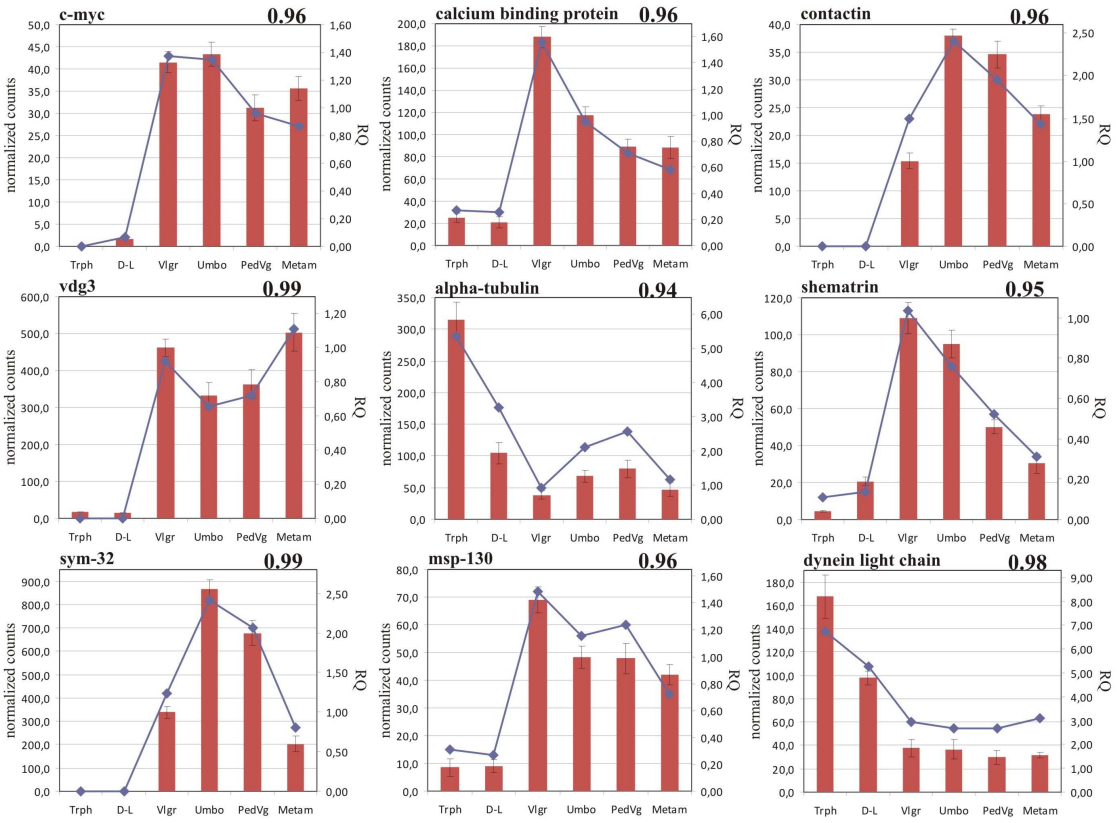

Figure 4.13 : Validations of differentially expressed genes defined by reads counting using qRT-PCR. In each graph the blue line describes the reads counting profile ad it is refers to left axis while red boxes describe qRT-PCR results and refer to right axis (95% confidence intervals are reported). The Pearson correlation coefficient has been reported upon each graph.

Therefore, we have repeated the annotation process using Blast-X *versus* the NCBI non redundant database with a cut-off of $<10^{-6}$. This second annotation process allowed to annotate only the 31,2% of differentially expressed genes, and assigned a GO description to only the 26,6% of transcripts.

In Table 4.5 we have reported the most represented terms of biological processes. As expected, many differentially expressed genes were assigned to GO terms about development and growth such as “anatomical structure development” (9.5%), “multicellular organismal development” (4%), and “regulation of biological process” (8%). However, the most abundant GO term is “macromolecule metabolic process” (20%), including pathways relative to protein and nucleic acid metabolism, with a peak of representation at umbo and pediveliger stage. An high number of differentially expressed genes are grouped in “energy process” (4%) GO term in umbo e pediveliger stages because in the last part of development the organism undergoes a very fats growth that is a very expensive process from an energetic point of view. A lot of sequences are grouped in GO term “immune related process” (8%) suggesting an

important role of defence system during planktonic life. We have decided to create new GO terms named “biomineralization process” and “byssus secretion process” to group all the genes involved in shell and byssus synthesis. Biological comments about GO results as well as the analysis of most significant expression signatures are reported in discussion chapter (Chapter 5).

Most relevant Gene Ontology terms	Stages			
	Early Stages	Veliger Stage	Umbo and Pediveliger Stages	Metamorphosis Stage
macromolecule metabolic process	6	18	63	43
anatomical structure development	7	18	19	19
immune related process	2	13	15	22
regulation of biological process	13	16	12	11
biomineralization process	9	14	10	8
primary metabolic process	5	14	13	7
cellular metabolic process	4	7	11	14
establishment of localization	8	8	7	3
multicellular organismal development	3	12	5	5
energy process	1	4	13	7
microtubule-based process	8	1	2	2
byssus secretion process	0	0	5	8

Table 4.5 : Most relevant Gene Ontology terms for Biological process category. For each Gene Ontology term we reported the number of referred differentially expressed genes for each stage. Terms are ordered from the most represented to the least. The last two terms are reported for their biological relevance.

4.11 Functional gene characterization

We decided to begin the functional characterization of tow annotated and validated differentially expressed genes.

Isogroup 306. Isogroup 306 has been annotated as the developmental regulated Vdg3, a protein involved in the development of digestive gland. Using primers designed for the qRT-PCR, we performed a 5’RACE and a 3’RACE to verify the correct assembling of 454 reads. Vgd3 has a transcript of 412 nucleotides with a coding sequence of 345 nt that results in a 115 amino acids long protein. Vdg3 is a tissue specific transcript and it is a good marker to define the digestive gland development. *In situ* hybridization of Vdg3 for each stage-specific larval stage was performed, results were reported in Figure 4.15.A. In agreement to reads counting profile (Figure 4.15.B) vdg3 starts to be expressed from the veliger stage, and it is interesting to note the central position of digestive gland during the larval development and its movement to the hinge during the total organ reorganization that occurs during mussel metamorphosis to allow the mantle cavity formation.

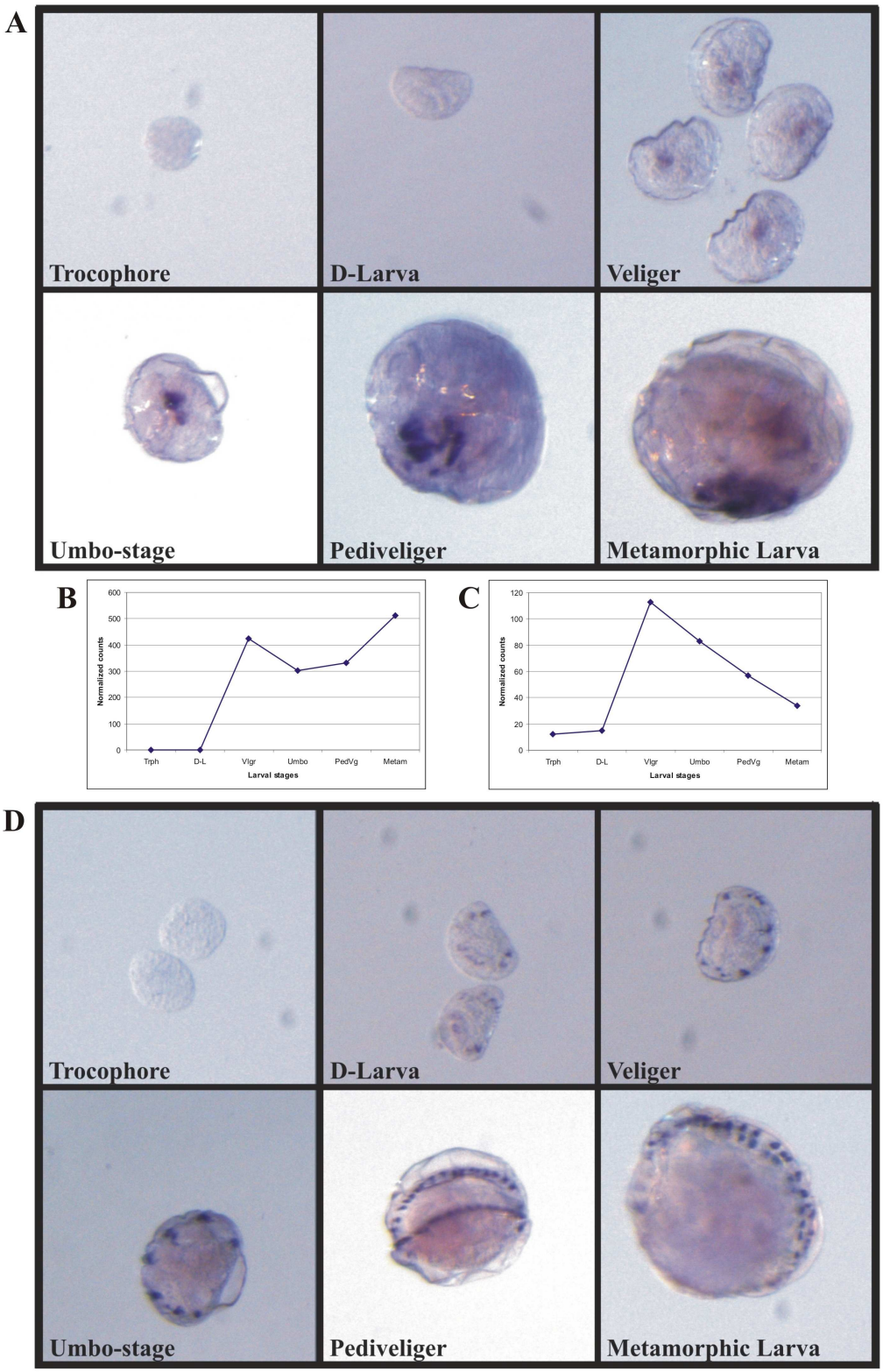

Figure 4.15 : *In situ* hybridization of two differentially expressed genes. A : *In situ* hybridization of vdg3. *In situ* hybridization was performed using an antisense probe of vdg3 (isogroup 306) hybridized on samples from each larval stage from trocophore to metamorphic larva. **B : Vdg3 reads counting profile. C : shematrin reads counting profile**. **D : *In situ* hybridization of shematrin.** *In situ* hybridization was performed using an antisense probe of shematrin (isogroup 306) hybridized on samples from each larval stage from trocophore to metamorphic larva.

.Isogroup 2203. Isogroup 2203 has been annotated as shematrin 4 that is one of the proteins responsible for shell formation. Its coding sequence could be 918 nt long according to *Pictada fucata.* Using primers designed for the qRT-PCR, we performed a 5'RACE and a 3'RACE to verify the correct assembling of 454 reads. Isogroup 2203 codify for 166 amino acids long protein, with a 659 nucleotides long transcript and a coding sequence of 498 nt. So our isogroup 2203 was shorter respect to the protein predicted from *Pictada fucata.* Shematrin 4 is member of a family of highly variable proteins that share a double domain extremely rich in glycine and a N-terminal signal peptide for their secretion. Our sequence showed these domains and it could be the first member of the Shemarin family discovered in *M. galloprovincialis*. We have performed *in situ* hybridization of isogroup2203 to determine its involvement in shell formation. As you can see in Figure 4.15.D signal of this transcript is localized in cells of the mantle edge thet are involved in shell secretion.This result is the first evidence of shematrin involvement in shell formation of Mediterranean mussel.

4.12 Embryo library analysis

The sequencing of embryo stage library has identified many genes with expression levels hundreds of times higher than other stages. For example histone H1 is a thousand times more expressed than in other libraries, where it is characterized by a stable profile. We have hypothesized that some of these transcripts could be oocyte stored mRNAs (osRNA), synthesized during oocyte maturation and then used immediately after fecundation to support the constitutive transcription during the fast embryonic divisions. To test our hypothesis we compared the level of expression of 13 putative osRNAs in not fertilized eggs and in one hour embryos using qRT-PCR. As you can see in Figure 4.16 only histone acetyltransferase is a embryo-specific gene while the transcription of other genes is also supported by a maternal component. In particular the expression levels referred to p53 and dermatopontin are the same before and after fertilization.

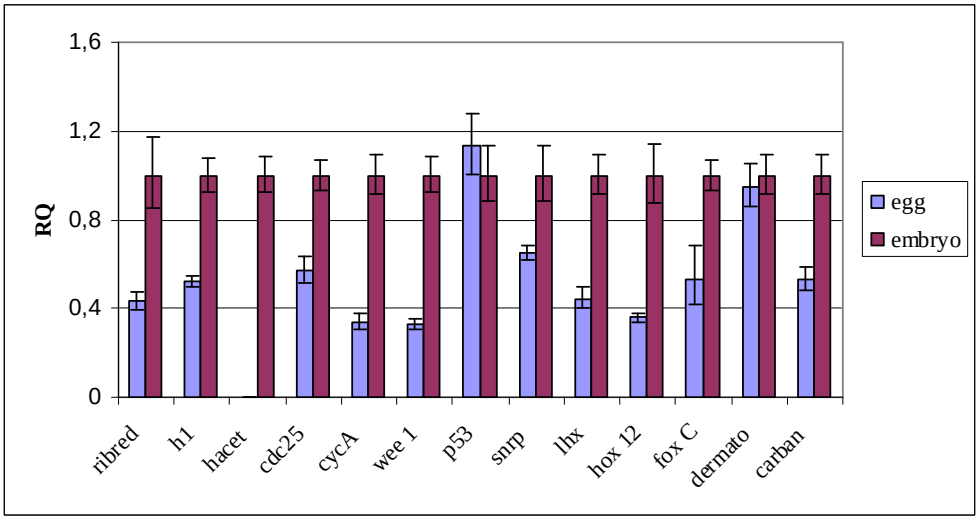

Figure 4.16 : Expression levels of putative oocyte stored mRNAs before and after fecundation. The expression levels of 13 putative osRNAs were obtained with qRT-PCR on not fertilized eggs (egg) and 1 hour embryons (embryo). Egg expression is referred to corresponding gene expression in embryo set as 1. 95% confidence intervals are reported. The analyzed genes are: ribonucleoside-diphosphate reductase (ribred); H1 histone member O oocyte-specific (h1), histone acetyltransferase type B (hacet); cdc25; cyclin-A (cycA); wee 1; p53; small chain small nuclear ribonucleoprotein polypeptide D3 (snrp); LIM class homeodomain transcription factor (lhx); hox12; foxC; dermatopontin 2 (dermato); carbonic anhydrase (carban).

5 Discussion

To elucidate the molecular mechanisms involved in the process of development from larva to the adult in *Mytilus galloprovincialis*, we have collected samples from seven developmental stages: from embryos to metamorphic larva. Then we have set up seven stage specific 3'-end cDNA libraries. That were sequenced with high-throughput next generation sequencing technologies. We were able to obtain: 10,200 consensus sequences of transcripts involved in larval development and to define their expression profiles calculating the 3'-EST frequency at each stage. The genome size of *Mytilus galloprovincialis* is about 1.6 Gb and Craft *et al.* (Craft JA, 2010) have proposed about 15,000 coding-protein genes, assuming that is not characterized by an high proportion of repeat sequences. On the basis of this hypothesis, we were able to define the expression profiling of the 65% of genes. Since our study is focused on larval transcripts and the most of sequences were obtained from larval samples, this is a good percentage of transcription representation.

We have performed a Gene Ontology (GO) analysis of stage-specific transcripts to increase our knowledge about larval development of *Mytilus galloprovincialis*. Here we reported several observations about the most involved genes in developmental process (Appendix 1).

Control of development. Several isogroups were resulted homologous to the most common genes involved in developmental and growth processes. Some of these were over-expressed during early larval stages such as the homeobox protein Hox-11 or the histone-binding protein caf-1 (chromatin assembly factor 1) that is a chromatin repressor. In *Drosophila* caf-1 ortholog is characterized by high level of expression during embryogenesis and then decreases in larval stages (Bulger M, 1995). Instead other transcripts were preferentially expressed during the late stages such as notch that is one of the master genes involved in developmental process, or the transcriptional factor c-myc, implicated in the control of many cellular processes from cell growth to apoptosis, that is expressed during the entire larval period but it is over-expressed during veliger and late stages. A similar expression trend was observed for several epidermal growth factors or for the dickkopf protein that playing an important role in development inhibiting Wnt (Wingless-type MMTV integration site family) regulated processes (Chien AJ, 2009). Unfortunately, it is impossible to evaluate the real impact

of these genes on the developmental process because they regulate a lot of genes in different biological process, and so specific functional studies should be conducted.

Soma ferritin, cyclophilin A, and HSP-90 (heat shock protein) were up regulated in later larval stages in agreement to previous studies (Williams EA, 2009; Heyland A, 2011) that have demonstrated the involvement of these genes in the attainment and maintenance of the metamorphic states in marine and terrestrial mollusc, but their roles are not yet completely clear.

Nervous system development. Morphological observations demonstrate that neurogenesis begins during trocophore stage starting from two precursor cells that during D-larva stage are differentiated into neurons, and they form cerebral ganglia at pediveliger stage (Voronezhskaya EE, 2008). Analyzing our data, we have divided genes involved in neurogenesis in two groups: genes overexpressed at early stages and on the other hand transcripts up-regulated from veliger to metamorphic larva. The first group included elav 2 (embryonic lethal, abnormal vision), that is involved in the maturation of neurons from their immature state to the final differentiated state (Robinow S, 1989), and pleiotrophin a secreted growth factor that induces neurite outgrowth (Milner PG, 1992).The expression of these genes at trocophore stage is consistent with an early neurogenesis. Whereas, the second group included genes responsible for axonal growth such as Slit 1 and tenascin c (Yu YM, 2011); but also genes involved in the development of central nervous system (CNS), such as sco-spondin that modulates neuronal aggregation during the formation of CNS; or CNS fatty acid binding proteins that are critical structural components for normal development of CNS. Combining morphological and expression data, we can conclude that neurons are maintained separated during the first phase of neurogenesis, while the ganglia reorganization starts between D-larva and veliger stage.

A functioning nervous system is fundamental during the entire mussel larval life; during the metamorphosis transition - that is initiated by specific sensory signal - as well as during D-larva stage, where it is needed to control velum muscle. Analyzing expression data, isogroup00064, annotated as neuronal acetylcholine receptor, that plays a key role during the development of neuromuscular junction in early developmental stages of many organisms, showed maximum expression levels during early larval stage demonstrating its importance to develop a functioning velum as soon as possible to fed and swim. The formation of acetylcholine receptor is induced by agrin protein released from the motor axon during innervation. Surprisingly, agrin expression profiling as well

as other genes involved in formation and stabilization of neuromuscular junctions, such as kyphoscoliosis peptidase, showed high expression levels only during veliger and late stage when the neuronal and muscular development are greater. Maybe the neuromuscular junctions formation during early larval stages is mediated by other specific genes.In agreement to previous studies performed on mollusks (Heyland A, 2011) ependymin and apolipoprotein-D showed high expression levels even if their involvement in neurogenesis is not yet clear.

Muscle development. The muscle system of adult mussel is characterized of smooth muscles whereas larval musculature is more complex and consists of both striated and smooth muscles (Odintsova NA, 2007). Dyachuk and collegues (Dyachuk V, 2009) have observed the appearance of the first muscle cells at the trocophore stage forming the muscle ring in early D-larva stage. Muscle ring consists of striated fibers and controls velar movements. During veliger stage, the muscle ring is converted to three striated retractor muscles and, at the same time, begins the formation of adult smooth adductor muscles. Finally, striated pattern of mussel muscle filaments disappears during metamorphosis; this degeneration process ends with the destruction of velar muscles together with the velum. We have sequenced all the main components of the muscle filament sliding process, including several variants of myosin heavy and light chain, paramyosin (specific of invertebrate muscle) and tropomyosin, and we have not observed variations of their expression levels across the larval stages. The expression of these protein during the entire mussel larval period was previously observed by Dyachuk *et al*. and confirms the continuous muscle development during larval life. Isogroup01202, annotated as myosin light chain, is up-regulated at early stages showing the same expression trend of isogroup00312, annotated as desmin, that is a protein expressed in striated muscles. So, this myosin could be a specific isoform expressed during the formation of the first striated larval muscle. We also identified catchin (myrod) (isogroup02388), that is a myosin rod-like protein expressed by alternative splicing from a myosin heavy chain in the catch-muscles of bivalves. These observations are in agreement with our data: catchin is expressed from veliger stage during the development of smooth adult muscles. We have also identify twitchin, that is another gene involved in catch process, characterized by same expression levels across the stages. According to Matusovskii OS et *al*. (Matusovskii OS, 2010) inactive larval isoform of twitchin is expressed during mussel larval stages necessary for the correct contractile apparatus formation of molluscan smooth muscles.

Vascularisation. Isogroup00507 and isogroup01070 were respectively identified as angiopoietin 1 and vascular endothelial growth factor (VEGF). VEGF is active in angiogenesis by inducing endothelial cell proliferation, migration, and permeabilization of blood vessels. Angiopoietin 1 mediates blood vessel maturation and stability playing an important role in heart early development. According to our expression angiogenesis and heart development start from veliger stage.

Digestive apparatus development. The expression of vdg3, a marker of the digestive gland, demonstrated the development of digestive apparatus between D-larva and veliger stage. Expression levels of other genes involved in digestive process such as lipases, glucanases, and mefrin A are in agreement to vgd3. Our *in situ* hybridization data have demonstrated no vgd3 signal in D-larva stage. *M. galloprovincialis* larvae start to feed immediately after the attainment of the D-larva stage using a first simple digestive tract. Since functioning of mussel digestive system is based on pinocitois, feeding could start before the complete development of digestive gland. Cathepsin L1 and L2 were over-expressed from veliger stage, whereas cathepsin B showed relatively high expression levels also in early stges. In bivalve, vitellin reserves stored in the oocyte are used in first two larval stages, then larvae become totally dependent on plankton for food. Wang *et al*. (Wang X, 2011) have demonstrated the role of cathepsin B in embryonic vitellin degradation in *M. meretrix*. Our data seems to confirm this observation. Moreover, these authors also suggested the involvement of this protein in apoptotic degradation of velum and velar striated muscles that occurs during metamorphosis. We were not able to confirm this observation in *M. galloprovincialis* because we have not detected up-regulation of chatepsins during the last larval stage.

Immune system and defense. We found an higher number of genes involved in immunity and defense processes. Huan and colleagues (Huan P, 2011) have demonstrated that the expression of genes involved in defense process during bivalves larval life is a physiological response to the planktonic life. Larvae, after D-larva stage development, start to take food from environment and they require an higher activity of the immune system. Analyzing expression data of genes involved in immune system and defense we have identified three category: genes expressed at early stages, from veliger stage, and after settlement. Only apextrin (Estévez-Calvar N, 2011) and a protein similar to immunoglobulin IgM heavy chain showed high expression levels in early stages. So few but specialized genes are involved in defense at early stages, but this number increasesed dramatically from veliger stage when larvae start to feed.

Elastase, defensin, c1q domain containing proteins, and several tipes of lectins were expressed from veliger stage. Some of these genes are proteases or their inhibitor attesting the close relationship between defense system and digestive apparatus that is the first organ exposed to the environment. Finally, after mussel settlement we observed an increased expression levels for the high variable antimicrobial peptides: myticin b, myticin c, and myticin d. We also identified 48 larval specific variants of myticin c.

Cilia and velum development. We have observed a high expression levels of alpha-tubulin and dynein at early larval stages. Other 6 genes of GO category named "microtubule based process" showed the same expression trend, such as tubulin polymerization-promoting protein 3 and Tektin 2 that play key roles in microtubules assembling; or the intraflagellar transport 20 that is involved in transport of particles during ciliary process assembly. Over-expression of these genes could be associated to the formation of the ciliary bands that are the first differentiating organ in mussel trocophore stage. The same genes were also over-expressed in D-larva stage to control development of velum (Damen WG, 1994). Our results are olso supported by the expression pattern of caveolin that is a marker gene for early specification of ciliary bands and velum as demonstrated by Arenas-Mena *et al.* (Arenas-Mena C, 2007) in early larva stages of *Hydroides elegans* with *in situ* hybridization.

Growth regulation. Between umbo-stage and pediveliger stage *M. galloprovincialis* undergoes a very fast growth that leads the organism to nearly double its size in few days. This event is associated with notably increase of over-expressed genes associated with macromolecule metabolic processes such as protein synthesis, folding and catabolism. We found 31 ribosomal proteins involved in protein synthesis and 5 proteasome subunits. Isogroup00825, annotated as T-complex 1 subunit zeta, is a molecular chaperone that assists protein folding. Isogroup02458, a peptidyl-prolyl cis-trans isomerase 5, is able to accelerate proteins folding. Protein synthesis and turnover are the first determinants of growth but require large amount of energy. In fact 13 over-expressed genes in late stages are grouped in the energy process GO category. We found genes of glycolysis (for example phosphoglycerate mutase and glyceraldehyde-3-phosphate dehydrogenase), of citric acid cycle (cytosolic malate dehydrogenase and isocitrate dehydrogenase), of fatty acid metabolism (hydroxyacyl dehydrogenase), and electron transport chain (nadh dehydrogenase subunit 3 and ATP synthase subunit a). Our results are confirmed by Meyer *et al.* (Meyer E, 2009) that have compared expression pattern of two population of *C. gigas* larvae characterized by growth

heterosis and they identified 34 marker genes associated to the “rapid growth” phenotype. We have found the 80% of transcripts identified by Meyer and we also discovered 50 new putative genes. Many of these genes showed over-expression also during metamorphosis that is considered another high energy consuming process.

Biomineralization. The most important genes in the biomineralization process are expressed constitutively during development. We have identify: carbonic anhydrase that provides HCO_3^- for calcification reaction (Ebanks SC, 2010), and dermatopontin, a high conserved component of extracellular matrix that produce mineralized tissues (Sarashina I, 2006). During larval period, mussels synthesize three different shell types characterized by peculiar chemical and physical properties. The determination of crystal type (calcite or aragonite) and the formation of the different layers (nacreous and prismatic) are regulated by protein composition of the organic matrix on which calcium carbonate ($CaCO_3$) is deposited. Genes involved in biomineralization process could be divided into three groups on the basis of expression signatures. Genes expressed at early stages are involved in the first larval shell formation; while over-expressed genes from veliger to pediveliger stages are involved in the second shell synthesis; and transcripts expressed from metamorphic stage contribute to the juvenile shell formation. The first larval shell is secreted during early stage and from an evolutional point of view it is one of the oldest mineralized tissue; it consists of a simple layer of amorfous $CaCO_3$ deposited on a glycine rich matrix similar to silk, egg case, egg chorion, and insect cuticle (Miyazaki Y, 2010). Several genes, specifically over-expressed at early stages, are grouped in glycine-rich family, such as isogroups 00050 and 00790 that showed omology with an eggshell protein and an insect cuticular protein respectively. The second larval shell is characterized by a double layer structure. The inner calcitic layer (nacreous layer) is similar to the first shell but organized into parallel sheets; this organization is regulated by lustrin A (Wustman BA, 2003) that, according to morphological observations, is up-regulated starting from veliger stage. Shematrin family (Yano M, 2006) is characterized by glycine-rich proteins that are mainly involved in the second shell layer (prismatic layer) formation in the adult. In our sequencing data we found for the first time in *M. galloprovincialis* several putative shematrin family members expressed during late larval life and characterized by shell type specificity. Galaxin/pearlin is a widely distributed calcium binding protein, considered essential in the biomineralization process of many mollusks (Reyes-Bermudez A, 2009). However, according to our data, it is expressed only from veliger

stage. Maybe, during early stage the Ca^{2+} is supplied by specific calcium binding proteins. We identified seven calcium binding proteins not directly involved in other biological processes (such as signaling or development pathways) and 4 of these proteins are specifically over-expressed at early stage. *In situ* hybridization as well as functional characterization of these proteins might help to confirm their putative role in biomineralization. The growth of two shell layers is synchronized and controlled by the activity of negative regulators of biomineralization such as perlwapin (Treccani L, 2006) and N19 (Yano M, 2007) that inhibit the crystallization of $CaCO_3$. According to our results, perlwapin is highly expressed starting from second shell secretion, whereas N19 is involved only in adult shell formation after metamorphosis. Ferritin is reported to play an important role in biomineralization (Fang D, 2011), because iron is the main metal component of the adult shell, but it was over-expressed starting from veliger stage demonstrating its involvement also in the second larval shell formation. Finally, chitinases, despite their role in immune and digestive processes, together with chitin synthetases are essential for remodeling chitin-containing structures, such as shell, during growth and development (Berdariotti F, 2007). Isogroup01047, similar to *C. gigas* chitinase 3, showed two peaks of expression at veliger and metamorphic larva stage when shell changes its mineral composition (Furuhashi T, 2009).

Byssus secretion. Among the primary components of byssus we have found all the three types of collagen (pre-collagen P, pre-collagen D, and pre-collagen NG) as well as 3 of the six foot proteinsshowing the same expression trend. Larva starts to synthesize the bissus from the pediveliger stage when the mussel is ready for the settlement. The synthesis of all the three types of collagen, responsible for the modulation of physical property of byssus, demonstrated that structural complexity of adult byssus is the same in larvae. We have not sequenced foot proteins involved in stem structure (foot proteins 1 and 2) and in plaque adhesion (foot protein 3) because these transcripts have low expression levels or in larval stages are expressed only larval specific isoforms.

Oocyte stored mRNA. After fertilization, eggs undergo a very rapid growth by alternating DNA synthesis and mitosis without noticeable G1 or G2 phases. During these fast divisions, cells have few time to transcribe mRNA even if Slater *et al.* (Slater DW, 1972) have observed that in sea urchin the rate of protein synthesis increases by 10–30 fold after fertilization. Therefore new proteins that are necessary to support this massive growth phase are obtained also from the stored maternal mRNAs. Oocyte makes, and stores in a dormant state, transcripts codifing for proteins essential during

the first embryo development. These genes are responsible for regulation of the timing of early cell division and for the massive production of chromatin, membranes, and cytoskeletal components as well as proteins (Radford HE, 2008). In fact, we found ribonucleoside reductases and histones involved in DNA synthesis and organization; cyclins involved in cell division regulation; tubulin essential component of the mitotic spindle; and genes involved in morphogenesis such as hox and fox. We also found dermatopontin and carbonic anidrase, that are involved in the secretion of the first larval shell demonstrating the beginning of the biomineralisation process at early stages of embryogenesis and attesting the importance of shell to guarantee protection to the organism as soon as possible. A list of the most relevant over-expressed genes and oocyte stored mRNA is reported in Appendix 2.

6 Conclusions

This work represents the first comprehensive analysis of gene expression profiles during the larval development in *Mytilus galloprovincialis.* We set up the 3'-end cDNA library protocol useful for next-generation sequencing obtaining thousands of new transcript sequences associated to reliable direct quantification of expression levels. We sequenced samples of the seven larval stages of *M. galloprovincialis*, from embryo to metamorphic larva, collected during a mating experiment. According to the peculiar transcriptional pattern associated to each stage, we defined the most relevant biological processes involved in the dynamic changes underlying larval development and their relative timing. We identified many differentially expressed genes involved in several key processes such as organ development, shell secretion, growth control, early mobility, and byssus secretion. Moreover, analyzing the expression pattern associated to embryo stage, we identified several oocyte stored mRNA including some transcripts involved in the secretion of the first larval shell. This study provides important basis for future research. It is important to note that a thousand of differentially expressed transcripts are not yet annotated and so their characterization will be important to better understand mechanism of mussel development. Now, we are starting the functional characterization of some differentially expressed unknown transcripts defining the full-length sequence and the localization with whole mount *in situ* hybridization.

7 References

Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J.J., Mayer, P. & Kawashima, E. 2000, "Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms", *Nucleic acids research*, vol. 28, no. 20, pp. E87.

Aji, L.P. 2011, "Spawning induction in Bivalves", *Journal penelitian sains*, vol. 14, no. 2, pp. 14207-14233.

Arenas-Mena, C., Wong, K.S. & Arandi-Forosani, N. 2007, "Ciliary band gene expression patterns in the embryo and trochophore larva of an indirectly developing polychaete", *Gene expression patterns : GEP*, vol. 7, no. 5, pp. 544-549.

Badariotti, F., Thuau, R., Lelong, C., Dubos, M.P. & Favrel, P. 2007, "Characterization of an atypical family 18 chitinase from the oyster Crassostrea gigas: evidence for a role in early development and immunity", *Developmental and comparative immunology*, vol. 31, no. 6, pp. 559-570.

Bardales, J.R., Cascallana, J.L. & Villamarin, A. 2011, "Differential distribution of cAMP-dependent protein kinase isoforms in various tissues of the bivalve mollusc Mytilus galloprovincialis", *Acta Histochemica*, vol. 113, no. 7, pp. 743-748.

Bennet-Clark, H.C. 1976, "MARINE MUSSELS: THEIR ECOLOGY AND PHYSIOLOGY. International Biological Programme 10 Edited by B. L. Bayne, Cambridge University Press, 1976. Pp. xvii+506. £22.00", *Experimental physiology*, vol. 61, no. 4, pp. 359-359.

Bulger, M., Ito, T., Kamakaka, R.T. & Kadonaga, J.T. 1995, "Assembly of regularly spaced nucleosome arrays by Drosophila chromatin assembly factor 1 and a 56-kDa histone-binding protein", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 25, pp. 11726-11730.

Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. 2010, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments", *BMC bioinformatics*, vol. 11, pp. 94.

Cannuel, R., Beninger, P.G., McCombie, H. & Boudry, P. 2009, "Gill Development and its functional and evolutionary implications in the blue mussel Mytilus edulis (Bivalvia: Mytilidae)", *The Biological bulletin*, vol. 217, no. 2, pp. 173-188.

Carl, C., Poole, A.J., Vucko, M.J., Williams, M.R., Whalan, S. & de Nys, R. 2011, "Optimising settlement assays of pediveligers and plantigrades of Mytilus galloprovincialis", *Biofouling*, vol. 27, no. 8, pp. 859-868.

Chien, A.J., Conrad, W.H. & Moon, R.T. 2009, "A Wnt survival guide: from flies to human disease", *The Journal of investigative dermatology*, vol. 129, no. 7, pp. 1614-1627.

Cloonan, N. & Grimmond, S.M. 2008, "Transcriptome content and dynamics at single-nucleotide resolution", *Genome biology*, vol. 9, no. 9, pp. 234.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. & Robles, M. 2005, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research", *Bioinformatics (Oxford, England)*, vol. 21, no. 18, pp. 3674-3676.

Craft, J.A., Gilbert, J.A., Temperton, B., Dempsey, K.E., Ashelford, K., Tiwari, B., Hutchinson, T.H. & Chipman, J.K. 2010, "Pyrosequencing of Mytilus galloprovincialis cDNAs: tissue-specific expression patterns", *PloS one*, vol. 5, no. 1, pp. e8875.

Dalloul, R.A., Long, J.A., Zimin, A.V., Aslam, L., Beal, K., Blomberg Le, A., Bouffard, P., Burt, D.W., Crasta, O., Crooijmans, R.P., Cooper, K., Coulombe, R.A., De, S., Delany, M.E., Dodgson, J.B., Dong, J.J., Evans, C., Frederickson, K.M., Flicek, P., Florea, L., Folkerts, O., Groenen, M.A., Harkins, T.T., Herrero, J., Hoffmann, S., Megens, H.J., Jiang, A., de Jong, P., Kaiser, P., Kim, H., Kim, K.W., Kim, S., Langenberger, D., Lee, M.K., Lee, T., Mane, S., Marcais, G., Marz, M., McElroy, A.P., Modise, T., Nefedov, M., Notredame, C., Paton, I.R., Payne, W.S., Pertea, G., Prickett, D., Puiu, D., Qioa, D., Raineri, E., Ruffier, M., Salzberg, S.L., Schatz, M.C., Scheuring, C., Schmidt, C.J., Schroeder, S., Searle, S.M., Smith, E.J., Smith, J., Sonstegard, T.S., Stadler, P.F.,

Tafer, H., Tu, Z.J., Van Tassell, C.P., Vilella, A.J., Williams, K.P., Yorke, J.A., Zhang, L., Zhang, H.B., Zhang, X., Zhang, Y. & Reed, K.M. 2010, "Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis", *PLoS biology*, vol. 8, no. 9, pp. e1000475.

Damen, W.G., van Grunsven, L.A. & van Loon, A.E. 1994, "Transcriptional regulation of tubulin gene expression in differentiating trochoblasts during early development of Patella vulgata", *Development (Cambridge, England)*, vol. 120, no. 10, pp. 2835-2845.

De Schrijver, J.M., De Leeneer, K., Lefever, S., Sabbe, N., Pattyn, F., Van Nieuwerburgh, F., Coucke, P., Deforce, D., Vandesompele, J., Bekaert, S., Hellemans, J. & Van Criekinge, W. 2010, "Analysing 454 amplicon resequencing experiments using the modular and database oriented Variant Identification Pipeline", *BMC bioinformatics*, vol. 11, pp. 269.

Dressman, D., Yan, H., Traverso, G., Kinzler, K.W. & Vogelstein, B. 2003, "Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 15, pp. 8817-8822.

Dyachuk, V. & Odintsova, N. 2009, "Development of the larval muscle system in the mussel Mytilus trossulus (Mollusca, Bivalvia)", *Development, growth & differentiation*, vol. 51, no. 2, pp. 69-79.

Ebanks, S.C., O'Donnell, M.J. & Grosell, M. 2010, "Characterization of mechanisms for Ca2+ and HCO3(-)/CO3(2-) acquisition for shell formation in embryos of the freshwater common pond snail Lymnaea stagnalis", *The Journal of experimental biology*, vol. 213, no. Pt 23, pp. 4092-4098.

Eo, S.H., Doyle, J.M., Hale, M.C., Marra, N.J., Ruhl, J.D. & Dewoody, J.A. 2012, "Comparative transcriptomics and gene expression in larval tiger salamander (Ambystoma tigrinum) gill and lung tissues as revealed by pyrosequencing", *Gene*, vol. 492, no. 2, pp. 329-338.

Estevez-Calvar, N., Romero, A., Figueras, A. & Novoa, B. 2011, "Involvement of pore-forming molecules in immune defense and development of the Mediterranean mussel (Mytilus galloprovincialis)", *Developmental and comparative immunology*, vol. 35, no. 10, pp. 1017-1031.

Eveland, & Eveland, A. 2008, "Transcript profiling by 3'-untranslated region sequencing resolves expression of gene families", *Plant Physiology*, vol. 146, no. 1, pp. 32.

Everett, M.V., Grau, E.D. & Seeb, J.E. 2011, "Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome", *Molecular ecology resources*, vol. 11 Suppl 1, pp. 93-108.

Fang, D., Xu, G., Hu, Y., Pan, C., Xie, L. & Zhang, R. 2011, "Identification of genes directly involved in shell formation and their functions in pearl oyster, Pinctada fucata", *PloS one*, vol. 6, no. 7, pp. e21860.

Faulhammer, D., Lipton, R.J. & Landweber, L.F. 2000, "Fidelity of enzymatic ligation for DNA computing", *Journal of computational biology : a journal of computational molecular cell biology*, vol. 7, no. 6, pp. 839-848.

Feng, B., Dong, L., Niu, D., Meng, S., Zhang, B., Liu, D., Hu, S. & Li, J. 2010, "Identification of immune genes of the Agamaki clam (Sinonovacula constricta) by sequencing and bioinformatic analysis of ESTs", *Marine biotechnology (New York, N.Y.)*, vol. 12, no. 3, pp. 282-291.

Francisco, C.J., Hermida, M.A. & Santos, M.J. 2010, "Parasites and symbionts from Mytilus galloprovincialis (Lamark, 1819) (Bivalves: Mytilidae) of the Aveiro Estuary Portugal", *The Journal of parasitology*, vol. 96, no. 1, pp. 200-205.

Frias-Lopez, J., Shi, Y., Tyson, G.W., Coleman, M.L., Schuster, S.C., Chisholm, S.W. & Delong, E.F. 2008, "Microbial community gene expression in ocean surface waters", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 10, pp. 3805-3810.

Furdek, M., Vahcic, M., Scancar, J., Milacic, R., Kniewald, G. & Mikac, N. 2012, "Organotin compounds in seawater and Mytilus galloprovincialis mussels along the Croatian Adriatic Coast", *Marine pollution bulletin*, .

Furuhashi, T., Schwarzinger, C., Miksik, I., Smrz, M. & Beran, A. 2009, "Molluscan shell evolution with review of shell calcification hypothesis", *Comparative biochemistry and physiology.Part B, Biochemistry & molecular biology*, vol. 154, no. 3, pp. 351-371.

Genovese, G., Faggio, C., Gugliandolo, C., Torre, A., Spano, A., Morabito, M. & Maugeri, T.L. 2012, "In vitro evaluation of antibacterial activity of Asparagopsis taxiformis from the Straits of Messina against pathogens relevant in aquaculture", *Marine environmental research,* vol. 73, no. 1, pp. 1-6.

Gérard, K., Bierne, N., Borsa, P., Chenuil, A. & Féral, J. 2008, "Pleistocene separation of mitochondrial lineages of Mytilus spp. mussels from Northern and Southern Hemispheres and strong genetic differentiation among southern populations", *Molecular phylogenetics and evolution,* vol. 49, no. 1, pp. 84-91.

Grossmann, V., Kohlmann, A., Klein, H.U., Schindela, S., Schnittger, S., Dicker, F., Dugas, M., Kern, W., Haferlach, T. & Haferlach, C. 2011, "Targeted next-generation sequencing detects point mutations, insertions, deletions and balanced chromosomal rearrangements as well as identifies novel leukemia-specific fusion genes in a single procedure", *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K,* vol. 25, no. 4, pp. 671-680.

Hestand, M.S., Klingenhoff, A., Scherf, M., Ariyurek, Y., Ramos, Y., van Workum, W., Suzuki, M., Werner, T., van Ommen, G.J., den Dunnen, J.T., Harbers, M. & 't Hoen, P.A. 2010, "Tissue-specific transcript annotation and expression profiling with complementary next-generation sequencing technologies", *Nucleic acids research,* vol. 38, no. 16, pp. e165.

Heyland, A., Vue, Z., Voolstra, C.R., Medina, M. & Moroz, L.L. 2011, "Developmental transcriptome of Aplysia californica", *Journal of experimental zoology.Part B, Molecular and developmental evolution,* vol. 316B, no. 2, pp. 113-134.

Huan, P., Wang, H. & Liu, B. 2012, "Transcriptomic Analysis of the Clam Meretrix meretrix on Different Larval Stages", *Marine biotechnology (New York, N.Y.),* vol. 14, no. 1, pp. 69-78.

Kinoshita, S., Wang, N., Inoue, H., Maeyama, K., Okamoto, K., Nagai, K., Kondo, H., Hirono, I., Asakawa, S. & Watabe, S. 2011, "Deep sequencing of ESTs from nacreous and prismatic layer producing tissues and a screen for novel shell formation-related genes in the pearl oyster", *PloS one,* vol. 6, no. 6, pp. e21238.

Lee, B.P., Messersmith, P.B., Israelachvili, J.N. & Waite, J.H. 2011, "Mussel-Inspired Adhesives and Coatings", *Annual review of materials research,* vol. 41, pp. 99-132.

Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G. & Webb, W.W. 2003, "Zero-mode waveguides for single-molecule analysis at high concentrations", *Science (New York, N.Y.),* vol. 299, no. 5607, pp. 682-686.

Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A. & Regev, A. 2010, "Comprehensive comparative analysis of strand-specific RNA sequencing methods", *Nature methods,* vol. 7, no. 9, pp. 709-715.

Lipson, D., Raz, T., Kieu, A., Jones, D.R., Giladi, E., Thayer, E., Thompson, J.F., Letovsky, S., Milos, P. & Causey, M. 2009, "Quantification of the yeast transcriptome by single-molecule sequencing", *Nature biotechnology,* vol. 27, no. 7, pp. 652-658.

Liu, S., Lin, L., Jiang, P., Wang, D. & Xing, Y. 2011, "A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species", *Nucleic acids research,* vol. 39, no. 2, pp. 578-588.

Livak, K.J. & Schmittgen, T.D. 2001, "Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method", *Methods (San Diego, Calif.),* vol. 25, no. 4, pp. 402-408.

Lon A., W. "Chapter 5 Neurobiology and behaviour of the scallop" in *Developments in Aquaculture and Fisheries Science* Elsevier, , pp. 317-356.

Longshaw, M., Feist, S.W. & Bateman, K.S. 2011, "Parasites and pathogens of the endosymbiotic pea crab (Pinnotheres pisum) from blue mussels (Mytilus edulis) in England", *Journal of invertebrate pathology,* .

Mahomed, W. & van den Berg, N. 2011, "EST sequencing and gene expression profiling of defence-related genes from Persea americana infected with Phytophthora cinnamomi", *BMC plant biology,* vol. 11, pp. 167.

Malone, J.H. & Oliver, B. 2011, "Microarrays, deep sequencing and the true measure of the transcriptome", *BMC biology,* vol. 9, pp. 34.

Marcheselli, M., Azzoni, P. & Mauri, M. 2011, "Novel antifouling agent-zinc pyrithione: stress induction and genotoxicity to the marine mussel Mytilus galloprovincialis", *Aquatic Toxicology (Amsterdam, Netherlands)*, vol. 102, no. 1-2, pp. 39-47.

Martin, J.A. & Wang, Z. 2011, "Next-generation transcriptome assembly", *Nature reviews.Genetics*, vol. 12, no. 10, pp. 671-682.

Matusovskii, O.S., Diachuk, V.A., Kiselev, K.V., Matusovskaia, G.G. & Shelud'ko, N.S. 2010, "Expression of several domains of twitchin and myorod in the ontogenesis of mussel Mytilus trossulus", *Biofizika*, vol. 55, no. 5, pp. 773-779.

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C., Zhang, Z., Ranade, S.S., Dimalanta, E.T., Hyland, F.C., Sokolsky, T.D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C.L., Li, B., Kotler, L., Stuart, J.R., Malek, J.A., Manning, J.M., Antipova, A.A., Perez, D.S., Moore, M.P., Hayashibara, K.C., Lyons, M.R., Beaudoin, R.E., Coleman, B.E., Laptewicz, M.W., Sannicandro, A.E., Rhodes, M.D., Gottimukkala, R.K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J.M., Eichler, E.E., Reese, M.G., De La Vega, F.M. & Blanchard, A.P. 2009, "Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding", *Genome research*, vol. 19, no. 9, pp. 1527-1541.

Meyer, E. & Manahan, D.T. 2010, "Gene expression profiling of genetically determined growth variation in bivalve larvae (Crassostrea gigas)", *The Journal of experimental biology*, vol. 213, no. 5, pp. 749-758.

Milner, P.G., Shah, D., Veile, R., Donis-Keller, H. & Kumar, B.V. 1992, "Cloning, nucleotide sequence, and chromosome localization of the human pleiotrophin gene", *Biochemistry*, vol. 31, no. 48, pp. 12023-12028.

Miyazaki, Y., Nishida, T., Aoki, H. & Samata, T. 2010, "Expression of genes responsible for biomineralization of Pinctada fucata during development", *Comparative biochemistry and physiology.Part B, Biochemistry & molecular biology*, vol. 155, no. 3, pp. 241-248.

Morga, B., Arzul, I., Faury, N. & Renault, T. 2010, "Identification of genes from flat oyster Ostrea edulis as suitable housekeeping genes for quantitative real time PCR", *Fish & shellfish immunology*, vol. 29, no. 6, pp. 937-945.

Morse, D.E., Duncan, H., Hooker, N. & Morse, A. 1977, "Hydrogen peroxide induces spawning in mollusks, with activation of prostaglandin endoperoxide synthetase", *Science (New York, N.Y.)*, vol. 196, no. 4287, pp. 298-300.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. 2008, "Mapping and quantifying mammalian transcriptomes by RNA-Seq", *Nature methods*, vol. 5, no. 7, pp. 621-628.

Ni, T., Corcoran, D.L., Rach, E.A., Song, S., Spana, E.P., Gao, Y., Ohler, U. & Zhu, J. 2010, "A paired-end sequencing strategy to map the complex landscape of transcription initiation", *Nature methods*, vol. 7, no. 7, pp. 521-527.

Obata, M., Sano, N. & Komaru, A. 2011, "Different transcriptional ratios of male and female transmitted mitochondrial DNA and tissue-specific expression patterns in the blue mussel, Mytilus galloprovincialis", *Development, growth & differentiation*, vol. 53, no. 7, pp. 878-886.

Odintsova, N.A., Diachuk, V.A. & Karpenko, A.A. 2007, "Development of the muscle system and contractile activity in the mussel Mytilus trossulus (Mollusca, Bivalvia)", *Ontogenez*, vol. 38, no. 3, pp. 235-240.

Ozsolak, F. & Milos, P.M. 2011, "Single-molecule direct RNA sequencing without cDNA synthesis", *Wiley interdisciplinary reviews.RNA*, vol. 2, no. 4, pp. 565-570.

Perocchi, F., Xu, Z., Clauder-Munster, S. & Steinmetz, L.M. 2007, "Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D", *Nucleic acids research*, vol. 35, no. 19, pp. e128.

Plessy, C., Bertin, N., Takahashi, H., Simone, R., Salimullah, M., Lassmann, T., Vitezic, M., Severin, J., Olivarius, S., Lazarevic, D., Hornig, N., Orlando, V., Bell, I., Gao, H., Dumais, J., Kapranov, P., Wang, H., Davis, C.A., Gingeras, T.R., Kawai, J., Daub, C.O., Hayashizaki, Y., Gustincich, S. & Carninci, P. 2010, "Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan", *Nature methods*, vol. 7, no. 7, pp. 528-534.

Radford, H.E., Meijer, H.A. & de Moor, C.H. 2008, "Translational control by cytoplasmic polyadenylation in Xenopus oocytes", *Biochimica et biophysica acta*, vol. 1779, no. 4, pp. 217-229.

Raftopoulou, E.K. & Dimitriadis, V.K. 2012, "Aspects of the digestive gland cells of the mussel Mytilus galloprovincialis, in relation to lysosomal enzymes, lipofuscin presence and shell size: Contribution in the assessment of marine pollution biomarkers", *Marine pollution bulletin*, .

Reyes-Bermudez, A., Lin, Z., Hayward, D.C., Miller, D.J. & Ball, E.E. 2009, "Differential expression of three galaxin-related genes during settlement and metamorphosis in the scleractinian coral Acropora millepora", *BMC evolutionary biology*, vol. 9, pp. 178.

Robinow, S., Campos, A.R., Yao, K.M. & White, K. 1988, "The elav gene product of Drosophila, required in neurons, has three RNP consensus motifs", *Science (New York, N.Y.)*, vol. 242, no. 4885, pp. 1570-1572.

Robinson, M.D., McCarthy, D.J. & Smyth, G.K. 2010, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data", *Bioinformatics (Oxford, England)*, vol. 26, no. 1, pp. 139-140.

Rosani, U., Varotto, L., Rossi, A., Roch, P., Novoa, B., Figueras, A., Pallavicini, A. & Venier, P. 2011, "Massively parallel amplicon sequencing reveals isotype-specific variability of antimicrobial peptide transcripts in Mytilus galloprovincialis", *PloS one*, vol. 6, no. 11, pp. e26680.

Rothberg, J.M. & Leamon, J.H. 2008, "The development and impact of 454 sequencing", *Nature biotechnology*, vol. 26, no. 10, pp. 1117-1124.

Saavedra, C. & Bachère, E. 2006, "Bivalve genomics", *Aquaculture*, vol. 256, no. 1–4, pp. 1-14.

Saeed, A.I., Bhagabati, N.K., Braisted, J.C., Liang, W., Sharov, V., Howe, E.A., Li, J., Thiagarajan, M., White, J.A. & Quackenbush, J. 2006, "TM4 microarray software suite", *Methods in enzymology*, vol. 411, pp. 134-193.

Sarashina, I., Yamaguchi, H., Haga, T., Iijima, M., Chiba, S. & Endo, K. 2006, "Molecular evolution and functionally important structures of molluscan Dermatopontin: implications for the origins of molluscan shell matrix proteins", *Journal of Molecular Evolution*, vol. 62, no. 3, pp. 307-318.

Siah, A., Dohoo, C., McKenna, P., Delaporte, M. & Berthe, F.C. 2008, "Selecting a set of housekeeping genes for quantitative real-time PCR in normal and tetraploid haemocytes of soft-shell clams, Mya arenaria", *Fish & shellfish immunology*, vol. 25, no. 3, pp. 202-207.

Siebert, S., Robinson, M.D., Tintori, S.C., Goetz, F., Helm, R.R., Smith, S.A., Shaner, N., Haddock, S.H. & Dunn, C.W. 2011, "Differential gene expression in the siphonophore Nanomia bijuga (Cnidaria) assessed with multiple next-generation sequencing workflows", *PloS one*, vol. 6, no. 7, pp. e22953.

Skibinski, D.O. & Beardmore, J.A. 1979, "A genetic study of intergradation between Mytilus edulis and Mytilus galloprovincialis", *Experientia*, vol. 35, no. 11, pp. 1442-1444.

Slater, D.W., Slater, I. & Gillespie, D. 1972, "Post-fertilization synthesis of polyadenylic acid in sea urchin embryos", *Nature*, vol. 240, no. 5380, pp. 333-337.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H. & Yaspo, M.L. 2008, "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome", *Science (New York, N.Y.)*, vol. 321, no. 5891, pp. 956-960.

Sun, C. & Waite, J.H. 2005, "Mapping chemical gradients within and along a fibrous structural tissue, mussel byssal threads", *The Journal of biological chemistry*, vol. 280, no. 47, pp. 39332-39336.

Treccani, L., Mann, K., Heinemann, F. & Fritz, M. 2006, "Perlwapin, an abalone nacre protein with three four-disulfide core (whey acidic protein) domains, inhibits the growth of calcium carbonate crystals", *Biophysical journal*, vol. 91, no. 7, pp. 2601-2608.

Venier, P., De Pitta, C., Bernante, F., Varotto, L., De Nardi, B., Bovo, G., Roch, P., Novoa, B., Figueras, A., Pallavicini, A. & Lanfranchi, G. 2009, "MytiBase: a knowledgebase of mussel (M. galloprovincialis) transcribed sequences", *BMC genomics*, vol. 10, pp. 72.

Venier, P., Varotto, L., Rosani, U., Millino, C., Celegato, B., Bernante, F., Lanfranchi, G., Novoa, B., Roch, P., Figueras, A. & Pallavicini, A. 2011, "Insights into the innate immunity of the Mediterranean mussel Mytilus galloprovincialis", *BMC genomics*, vol. 12, pp. 69.

Vera, M., Martinez, P., Poisa-Beiro, L., Figueras, A. & Novoa, B. 2011, "Genomic organization, molecular diversification, and evolution of antimicrobial peptide myticin-C genes in the mussel (Mytilus galloprovincialis)", *PloS one*, vol. 6, no. 8, pp. e24041.

Voronezhskaya, E.E., Glebov, K.I., Khabarova, M.Y., Ponimaskin, E.G. & Nezlin, L.P. 2008, "Adult-to-embryo chemical signaling in the regulation of larval development in trochophore animals: cellular and molecular mechanisms", *Acta Biologica Hungarica*, vol. 59 Suppl, pp. 117-122.

Voronezhskaya, E.E., Nezlin, L., Odintsova, N., Plummer, J. & Croll, R. 2008, *Neuronal development in larval mussel Mytilus trossulus (Mollusca: Bivalvia)*, Springer Berlin / Heidelberg.

Wang, X., Liu, B., Tang, B. & Xiang, J. 2011, "Potential role of cathepsin B in the embryonic and larval development of clam Meretrix meretrix", *Journal of experimental zoology.Part B, Molecular and developmental evolution*, vol. 316, no. 4, pp. 306-312.

Weiss, I.M. & Schonitzer, V. 2006, "The distribution of chitin in larval shells of the bivalve mollusk Mytilus galloprovincialis", *Journal of structural biology*, vol. 153, no. 3, pp. 264-277.

Wilkens, L.A. 1981, "Neurobiology of the Scallop. I. Starfish-Mediated Escape Behaviours", *Proceedings of the Royal Society of London.Series B, Biological Sciences*, vol. 211, no. 1184, pp. pp. 341-372.

Williams, E.A., Degnan, B.M., Gunter, H., Jackson, D.J., Woodcroft, B.J. & Degnan, S.M. 2009, "Widespread transcriptional changes pre-empt the critical pelagic-benthic transition in the vetigastropod Haliotis asinina", *Molecular ecology*, vol. 18, no. 5, pp. 1006-1025.

Wustman, B.A., Weaver, J.C., Morse, D.E. & Evans, J.S. 2003, "Structure-function studies of the lustrin A polyelectrolyte domains, RKSY and D4", *Connective tissue research*, vol. 44 Suppl 1, pp. 10-15.

Yano, M., Nagai, K., Morimoto, K. & Miyamoto, H. 2007, "A novel nacre protein N19 in the pearl oyster Pinctada fucata", *Biochemical and biophysical research communications*, vol. 362, no. 1, pp. 158-163.

Yano, M., Nagai, K., Morimoto, K. & Miyamoto, H. 2006, "Shematrin: a family of glycine-rich structural proteins in the shell of the pearl oyster Pinctada fucata", *Comparative biochemistry and physiology.Part B, Biochemistry & molecular biology*, vol. 144, no. 2, pp. 254-262.

Yu, Y.M., Cristofanilli, M., Valiveti, A., Ma, L., Yoo, M., Morellini, F. & Schachner, M. 2011, "The extracellular matrix glycoprotein tenascin-C promotes locomotor recovery after spinal cord injury in adult zebrafish", *Neuroscience*, vol. 183, pp. 238-250.

Zeng, V., Villanueva, K.E., Ewen-Campen, B.S., Alwes, F., Browne, W.E. & Extavour, C.G. 2011, "De novo assembly and characterization of a maternal and developmental transcriptome for the emerging model crustacean Parhyale hawaiensis", *BMC genomics*, vol. 12, no. 1, pp. 581.

Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. & Siebert, P.D. 2001, "Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction", *BioTechniques*, vol. 30, no. 4, pp. 892-897.

Zwaan, A. & Mathieu, M. 1992, *Cellular biochemistry and endocrinology*, NIOO, CEMO.

Appendix 1

Most relevant mRNA over-expressed during larval development. For each isogroup was reported the annotation and the normalized count across larval stages. mRNAs cited in Discussion chapter are listed in bold.

Control of development

Isogroup	Annotation	Stages					
		Trocophore	D-Larva	Veliger	Umbo-Stage	Pediveliger	Methamorphic Larva
01078	**homeobox protein Hox-11**	**25**	**15**	3	0	0	0
09989	**histone-binding protein caf1**	**12**	**7**	**4**	0	1	0
02393	**notch homolog Scalloped wings**	0	4	**17**	**14**	**12**	**14**
00074	**c-myc**	0	2	**43**	**42**	**30**	**27**
02180	**Multiple epidermal growth factor**	8	6	**82**	**40**	**58**	**67**
01296	**egf-like protein**	0	0	0	0	**11**	**27**
05927	**dickkopf protein**	25	18	**145**	**68**	**81**	**67**
03058	jerky protein homolog	9	10	2	0	0	1
00237	tetraspanin 7	1	2	**13**	**12**	**14**	**13**
01000	Cytokeratin-9	**51**	**33**	0	0	0	0
08121	cysteine-rich secretory protein LCCL domain containing 2	73	31	**194**	**67**	27	27
01069	**soma ferritin**	0	0	1	**26**	**18**	**22**
00210	**cyclophilin A**	23	29	**88**	**86**	**76**	**107**
00217	**HSP 90**	1	2	15	12	**26**	**78**

Nervous system development

Isogroup	Annotation	Stages					
		Trocophore	D-Larva	Veliger	Umbo-Stage	Pediveliger	Methamorphic Larva
07238	**elav 2**	**57**	**52**	31	24	22	16
00445	**pleiotrophin**	**42**	**44**	18	8	6	4
00777	cysteine-rich motor neuron	**90**	**98**	4	2	0	0
00227	cysteine-rich motor neuron	0	0	**15**	**15**	**18**	8
01760	cysteine-rich motor neuron	0	0	0	4	5	**25**
00196	**slit homolog 1**	12	11	**167**	**61**	**50**	**83**
08156	**tenascin c**	0	1	4	**6**	**15**	**9**
02271	**sco-spondin**	0	0	4	4	**6**	**15**
00019	**fatty acid binding protein brain**	21	33	**110**	**160**	**120**	**97**
00561	contactin associated 2	0	0	**24**	**34**	**26**	**27**

00517	contactin-associated 5-like	0	0	**23**	**37**	**30**	**22**
00064	**Neuronal acetylcholine receptor subunit alpha-7**	**193**	**183**	18	12	9	12
01237	**agrin**	15	22	**110**	**89**	**84**	**69**
01199	**kyphoscoliosis peptidase**	1	2	**9**	**15**	**14**	**21**
00235	**Ependymin**	76	54	**546**	**242**	**209**	**265**
00489	ependymin-related protein	6	4	**53**	**66**	**121**	**96**
00041	**apolipoprotein D**	2	5	**49**	**46**	**50**	**41**
01324	dopamine beta hydroxylase-like protein	0	4	**69**	**63**	**78**	**51**

Muscle development

Isogroup	Annotation	Stages					
		Trocophore	D-Larva	Veliger	Umbo-Stage	Pediveliger	Methamorphic Larva
01202	**myosin light chain**	**294**	**298**	55	59	65	7
00312	**desmin like**	**276**	**161**	23	18	19	12
02388	**catchin**	0	1	**16**	**20**	**22**	**20**

Vascularisation

Isogroup	Annotation	Stages					
		Trocophore	D-Larva	Veliger	Umbo-Stage	Pediveliger	Methamorphic Larva
00507	**angiopoietin-like 1**	0	0	**13**	**12**	**21**	**21**
01070	**vascular endothelial growth factor**	0	6	**18**	**16**	**11**	**20**

Digestive apparatus development

Isogroup	Annotation	Stages					
		Trocophore	D-Larva	Veliger	Umbo-Stage	Pediveliger	Methamorphic Larva
00306	**vdg3**	0	1	**421**	**302**	**327**	**516**
02057	gastric triacylglycerol lipase	0	0	**9**	**12**	**29**	**38**
02466	meprin A alpha	2	11	**146**	**62**	**57**	**54**
00896	pancreatic lipase-related protein	0	0	**15**	**6**	**11**	**7**
01025	endo-1,3-beta-D-glucanase	0	0	**88**	**104**	**141**	**96**
01194	endo-1,4-beta-D-glucanase	0	0	**16**	**13**	**31**	**21**
08895	lactase phlorizinhydrolase	0	0	**16**	**6**	**9**	**11**
00041	**cathepsin B**	54	79	**232**	**279**	**394**	**179**
00121	**cathepsin L1**	2	6	**129**	**105**	**105**	**53**

00866	**cathepsin L2**	4	5	**27**	**50**	**57**	**22**
00721	**caspase-3 precursor**	2	0	**3**	**4**	**3**	**15**

Immune system and defense

Isogroup	Annotation	Stages					
		Trocophore	D-Larva	Veliger	Umbo-Stage	Pediveliger	Methamorphic Larva
00184	**immunoglobulin IgM heavy chain like**	**308**	**405**	1	0	0	0
02596	**early apextrin**	**165**	**115**	16	5	0	1
00711	**defensin precursor**	0	0	**20**	**17**	**11**	**9**
01168	**Elastase**	0	0	**54**	**31**	**25**	**32**
02839	**c1q domain containing protein**	0	0	**9**	**11**	**13**	**7**
00324	**c1q domain containing protein**	0	0	**8**	**4**	**9**	**14**
01142	similar to H.sapiens MPEG1, macrophage expressed gene 1	0	0	**299**	**161**	**131**	**85**
00171	C-type Lectin	39	40	**135**	**163**	**116**	**112**
02225	C-type Lectin	0	0	**7**	**4**	**16**	**16**
02024	alpha-n-acetylgalactosamine-binding lectin	0	0	**10**	**7**	**12**	**14**
00775	salivary c-type lectin	0	1	**13**	**4**	**36**	**37**
03010	c-type lectin 5	0	0	**5**	**4**	**2**	**10**
09299	galactose-specific c-type	0	0	0	1	3	**12**
00372	sialic acid binding lectin	0	0	0	0	1	**9**
03435	late apextrin	0	0	0	0	0	**15**
01419	**myticin b**	0	0	0	2	9	**115**
05789	**mytilin c**	0	0	0	2	1	**58**
05043	**mytilin d**	0	0	0	0	1	**20**
05239	complement c1q-like protein 2	0	0	0	0	0	**19**
02006	lysozyme	2	1	**12**	**21**	**10**	**22**
00617	lysozyme 2	1	1	**12**	**27**	**39**	**16**
00727	lysozyme g	0	0	**7**	**9**	**3**	**6**
04610	serine protease CFSP3	0	0	**46**	**24**	**12**	3
00196	Serine protease inhibitor Cvsi-2	12	11	**167**	**61**	**50**	**83**
00292	cysteine protease inhibitor	7	6	**56**	**61**	**98**	**61**
02075	hepatopancreas kazal-type proteinase inhibitor	10	8	**71**	**58**	**71**	**35**
01597	Kunitz domain-containing protein	6	6	**79**	**44**	**46**	**52**

Cilia and velum development

Isogroup	Annotation	Stages					
		Trocophore	D-Larva	Veliger	Umbo-Stage	Pediveliger	Methamorphic Larva
07025	**alpha-tubulin**	**258**	**176**	49	114	138	63
00967	**dynein heavy chain**	**62**	**48**	5	3	8	7
00033	**dynein light chain**	**138**	**108**	60	55	55	63
00114	**tubulin polymerization-promoting protein family member 3**	**415**	**298**	68	61	66	79
05145	**Tektin 2**	**74**	**94**	26	34	39	36
02082	**intraflagellar transport protein 20**	**24**	**20**	11	4	8	6
02433	growth arrest-specific 8	**62**	**55**	26	29	24	15
02504	ropporin 1-like	**55**	**69**	30	14	16	8
00523	**caveolin 3**	**88**	**92**	20	25	24	9

Growth regulation

Isogroup	Annotation	Stages					
		Trocophore	D-Larva	Veliger	Umbo-Stage	Pediveliger	Methamorphic Larva
00825	**T-complex protein 1 subunit zeta**	9	12	**20**	**37**	**29**	**21**
02458	**peptidyl-prolyl cis-trans isomerase 5**	0	2	4	**29**	**36**	7
00159	**phosphoglycerate mutase**	2	2	**11**	**25**	**25**	**12**
02380	**glyceraldehyde-3-phosphate dehydrogenase**	2	1	**14**	**24**	**19**	**11**
00254	**cytosolic malate dehydrogenase**	14	19	**45**	**47**	**55**	22
00703	**isocitrate dehydrogenase**	6	2	**13**	**31**	**30**	**13**
02002	**hydroxyacyl dehydrogenase**	5	4	**16**	**15**	**19**	11
00912	**nadh dehydrogenase subunit 1**	17	18	**37**	**44**	**35**	25
00314	nadh dehydrogenase subunit 2	29	32	**51**	**67**	**58**	41
02010	nadh dehydrogenase subunit 3	36	29	**57**	**81**	**84**	**45**
01665	**ATP synthase subunit a**	38	45	**70**	**113**	**70**	**64**

Byssus secretion

Isogroup	Annotation	Stages					
		Trocophore	D-Larva	Veliger	Umbo-Stage	Pediveliger	Methamorphic Larva
00568	foot protein 4 variant 1	0	0	0	0	**8**	**20**
01295	foot protein 4 variant 2	0	0	0	0	**3**	**18**
07333	foot protein 5	0	0	0	0	**1**	**8**
08018	foot protein 5 variant 3	0	0	0	0	**3**	**5**

09144	foot protein 6 variant 3	0	0	0	0	**1**	**16**
02406	pre-collagen D	0	0	0	0	**6**	**10**
02733	pre-collagen NG	0	0	0	0	**1**	**3**
06430	pre-collagen P	0	0	0	0	**2**	**15**

Biomineralization

Isogroup	Annotation	Stages					
		Trocophore	D-Larva	Veliger	Umbo-Stage	Pediveliger	Methamorphic Larva
01617	unknown glycine rich protein	**138**	**92**	0	0	0	0
00023	unknown glycine rich protein	**63**	**89**	0	0	0	0
09311	unknown glycine rich protein	**12**	**7**	0	0	0	0
00050	**Glycine -rich eggshell protein**	**655**	**751**	0	0	0	0
00790	**cuticular protein glycine-rich 1**	**21**	**27**	0	0	0	0
00279	chitin binding protein	**44**	17	13	3	6	5
02203	**shematrin like**	12	15	**113**	**83**	**57**	34
07414	**shematrin like**	0	0	1	4	10	**51**
02246	**lustrin A**	0	0	3	**13**	**11**	**11**
01234	**galaxin/pearlin**	0	0	**17**	**22**	**23**	**44**
01069	**ferritin**	0	1	**11**	**26**	**18**	**22**
00083	**perlwapin**	6	3	**54**	**22**	**32**	**20**
04112	**N19**	0	0	0	0	0	**16**
01047	**chitinase 3**	9	15	**106**	48	69	**104**
08732	mantle protein	**19**	**19**	4	1	0	1
00486	mantle protein 12 like	**81**	**57**	12	3	4	2
00134	mantle protein 9 like	24	29	**160**	**103**	**126**	57
09463	mantle protein 4 like	0	0	0	**9**	**5**	**7**
01163	calcium-binding protein	**12**	**10**	5	1	2	4
01714	calcium-binding protein	**18**	**6**	**6**	1	1	4
02798	calcium-binding protein	**89**	**98**	34	32	16	2
05175	secreted modular calcium-binding protein 1	**14**	**4**	0	0	0	1
04016	calcium binding protein	0	4	**17**	1	3	3
00664	calcium binding protein	0	0	30	24	**126**	**197**
04699	calmodulin putative	4	1	**30**	**30**	**41**	**12**

Appendix 2

Most relevant over-expressed genes at embryo stage. Validated oocyte stored mRNAs are listed in bold.

Isogroup	Annotation
00086	**H1 histone family, member O, oocyte-specific**
01422	histone acetyltransferase type B catalytic subunit
00382	**ribonucleoside-diphosphate reductase small chain**
00521	**small nuclear ribonucleoprotein polypeptide D3**
02365	**cyclin-A**
00240	cyclin-B
00821	alpha-tubulin
07298	**p53**
01722	**cdc 25**
04695	**wee 1 kinase**
06665	**foxC**
03452	**hox 12**
02559	**LIM class homeodomain transcription factor**
00195	**carbonic anhydrase**
00695	**dermatopontin 2**
00224	acetylcholine receptor alpha subunit
02887	lipopolysaccharide-induced TNF-alpha factor
00204	metallothionein-10B
01101	nucleolar protein-like
05735	RAB interacting factor-like
01132	TOM22
07423	C1q and tumor necrosis factor related protein 7-like
03853	SWI/SNF related, matrix associated, actin dependent regulator of chromatin
03810	splA/ryanodine receptor domain and SOCS box containing 3
03787	polymerase (RNA) II
01391	rnase H family member
00189	TIM23

Acknowledgements

Firstly, I would like to thank Prof. Gerolamo Lanfranchi for giving me an opportunity to work in his lab, and for his supervision and support.
I specifically thank Cristiano De Pittà for his constant guidance, encouragement, and discussions in day to day life in the lab.
I wish also to thank all those people who helped me a lot in the lab during these years, in particular Chiara Romualdi, Gabriele Sales, Alessandro Albiero, and Davide Risso for data analysis support and Enrico Moro for his assistance in *in situ* hybridization experiments.